IBM SPSS Modeler 19.0 Applications Guide

Nota

Antes de utilizar essas informações e o produto que elas suportam, leia as informações em "Avisos" na página 335.

Informações sobre o produto

Esta edição se aplica à versão 19, liberação 0, modificação 0 do IBM® SPSS Modelador e a todas as liberações e modificações subsequentes até que seja indicado de outra forma em novas edições.

© Copyright International Business Machines Corporation .
US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Capítulo 1. SobreIBM SPSS Modelador

O IBM SPSS Modelador é um conjunto de ferramentas de mineração de dados que permite desenvolver rapidamente modelos preditivos usando o conhecimento de negócios, e implementá-los em operações de negócios para melhorar a tomada de decisão. Projetado em torno do modelo CRISP-DM padrão de mercado, o IBM SPSS Modelador suporta todo o processo de mineração de dados, a partir dos dados para melhores resultados de negócios.

OIBM SPSS Modelador oferece uma variedade de métodos de modelagem tomados do aprendizado de máquina, inteligência artificial e estatística. Os métodos disponíveis na paleta Modelagem permitem derivar informações novas a partir dos dados, e desenvolver modelos preditivos. Cada método possui certas forças e é mais adequado para certos tipos de problemas.

O ModeladorSPSS pode ser comprado como um produto independente, ou usado como um cliente na combinação com o Servidor doSPSS Modeler. Várias opções adicionais também estão disponíveis, conforme resumidas nas seções a seguir. Para mais informações, consulte https://www.ibm.com/analytics/us/en/technology/spss/.

Produtos IBM SPSS Modelador

A família de produtos IBM SPSS Modelador e o software associado abrangem o seguinte.

- IBM SPSS Modelador
- ServidorIBM SPSS Modeler
- Console de administraçãoIBM SPSS Modeler (incluído com IBM SPSS Gerente de implantação)
- IBM SPSS Lote do modelador
- Editor de soluçõesIBM SPSS Modeler
- ServidorIBM SPSS Modeler adaptadores para IBM SPSS Serviços de colaboração e implantação

IBM SPSS Modelador

ModeladorSPSS é uma versão funcionalmente completa do produto que você instala e executa em seu computador pessoal. É possível executar o ModeladorSPSS no modo local como um produto independente ou usá-lo no modo distribuído com ServidorIBM SPSS Modeler para melhorar o desempenho em conjuntos de dados grandes.

Com o ModeladorSPSS, é possível construir modelos preditivos exatos de maneira rápida e intuitiva, sem programação. Usando a interface visual exclusiva, é possível visualizar facilmente o processo de mineração de dados. Com o suporte da análise avançada integrada ao produto, é possível descobrir tendências e padrões ocultos anteriormente em seus dados. É possível modelar resultados e entender os fatores que os influenciam, permitindo que você aproveite as vantagens das oportunidades de negócios e diminua os riscos.

ModeladorSPSS está disponível em duas edições: SPSS Modeler Professional e SPSS Modeler Premium. Consulte o tópico “Edições do IBM SPSS Modelador” na página 2 para obter mais informações.

ServidorIBM SPSS Modeler

ModeladorSPSS usa uma arquitetura de cliente/servidor para distribuir solicitações para operações cheias de recursos para poderosos softwares de servidor, resultando em desempenho mais rápido em conjuntos de dados maiores.

Servidor doSPSS Modeler é um produto licenciado separadamente que é executado de forma contínua no modo de análise distribuído em um host do servidor com uma ou mais instalações do IBM SPSS Modelador. Dessa maneira, o Servidor doSPSS Modeler fornece desempenho superior em conjuntos de dados grandes, pois operações com uso intensivo de memória podem ser executadas no servidor

sem fazer download dos dados no computador cliente. ServidorIBM SPSS Modeler também fornece suporte para otimização de SQL e capacidades de modelagem dentro da base de dados, entregando mais benefícios para o desempenho e a automação.

Console de administraçãoIBM SPSS Modeler

O Console de administração do Modeler é uma interface gráfica de usuário para o gerenciamento de muitas das opções de configuração Servidor doSPSS Modeler , que também são configuráveis por meio de um arquivo de opções. O console é incluído em IBM SPSS Gerente de implantação, pode ser usado para monitorar e configurar suas instalações Servidor doSPSS Modeler , e está disponível gratuitamente para os clientes atuais Servidor doSPSS Modeler . O aplicativo pode ser instalado somente em computadores Windows; no entanto, ele pode administrar um servidor instalado em qualquer plataforma suportada.

IBM SPSS Lote do modelador

Embora geralmente a mineração de dados seja um processo interativo, também é possível executar o ModeladorSPSS a partir de uma linha de comandos, sem a necessidade de uma interface gráfica com o usuário. Por exemplo, você pode ter tarefas repetidas ou de longa execução que deseja executar sem intervenção do usuário. SPSS Lote do modelador é uma versão especial do produto que fornece suporte para capacidades de análise completa do ModeladorSPSS sem acessar a interface com o usuário regular. Servidor doSPSS Modeler é necessário para usar o SPSS Lote do modelador.

Editor de soluçõesIBM SPSS Modeler

Editor de soluçõesSPSS Modeler é uma ferramenta que permite criar uma versão do pacote de um fluxo do ModeladorSPSS que pode ser executado por um mecanismo de tempo de execução externo ou integrado a um aplicativo externo. Dessa maneira, é possível publicar e implementar fluxos completos do ModeladorSPSS para uso em ambientes que não têm o ModeladorSPSS instalado. Editor de soluçõesSPSS Modeler é distribuído como parte do serviço IBM SPSS Serviços de colaboração e implantação - Pontuação, para o qual uma licença separada é necessária. Com essa licença, você recebe o Tempo de execuçãoSPSS Modeler Solution Publisher, que permite executar os fluxos publicados.

Para obter mais informações sobre Editor de soluçõesSPSS Modeler, consulte a documentação do IBM SPSS Serviços de colaboração e implantação. A documentação IBM SPSS Serviços de colaboração e implantação IBM contém seções chamadas " IBM SPSS Modeler Solution Publisher " e " IBM SPSS Analytics Toolkit"

ServidorIBM SPSS Modeler Adaptadores para IBM SPSS Serviços de colaboração e implantação

Inúmeros adaptadores para o IBM SPSS Serviços de colaboração e implantação estão disponíveis para permitir que o ModeladorSPSS e o Servidor doSPSS Modeler interajam com um repositório do IBM SPSS Serviços de colaboração e implantação. Dessa forma, um fluxo do ModeladorSPSS implementado no repositório pode ser compartilhado por diversos usuários ou acessado a partir do aplicativo thin client VantagensIBM SPSS Modeler. Você instala o adaptador no sistema que hospeda o repositório.

Edições do IBM SPSS Modelador

ModeladorSPSS está disponível nas seguintes edições.

SPSS Modeler Professional

SPSS Modeler Professional fornece todas as ferramentas necessárias para você trabalhar com a maioria dos tipos de dados estruturados, como comportamentos e interações controlados em sistemas CRM, demográficos, comportamento de compra e dados de vendas.

SPSS Modeler Premium

SPSS Modeler Premium é um produto licenciado separadamente que se estende SPSS Modeler Professional para trabalhar com dados especializados e com dados de texto não estruturados. SPSS Modeler Premium inclui IBM SPSS Modelador de análise de texto:

IBM SPSS Modelador de análise de texto usa tecnologias de linguística avançada e processamento de linguagem natural (NLP) para processar rapidamente uma grande variedade de dados de texto não estruturados, extrair e organizar conceitos chave e agrupar esses conceitos em categorias. Categorias e conceitos extraídos podem ser combinados com dados estruturados existentes, como demográficos, e aplicados à modelagem usando o conjunto completo de ferramentas de mineração de dados do IBM SPSS Modelador para gerar decisões melhores e mais focadas.

AssinaturaIBM SPSS Modeler

AssinaturaIBM SPSS Modeler fornece todas as mesmas capacidades de analítica preditiva que o cliente IBM SPSS Modelador tradicional. Com a edição de Assinaturas, é possível fazer o download de atualizações do produto regularmente.

Documentação

A documentação está disponível no menu **Ajuda** em ' ModeladorSPSS. Isso abre a documentação on-line IBM, que está sempre disponível fora do produto.

A documentação completa de cada produto (incluindo instruções de instalação) também está disponível em formato PDF. Consulte a seguinte página de suporte: **SPSS Modeler 19.0 documentation**.

Documentação do SPSS Modeler Professional

O conjunto de documentações do SPSS Modeler Professional (excluindo instruções de instalação) é o seguinte.

- **IBM SPSS Modelador User's Guide.** Introdução geral para usar ModeladorSPSS, incluindo como construir fluxos de dados, manipular valores ausentes, construir expressões CLEM , trabalhar com projetos e relatórios, e streams de pacotes para implementação em IBM SPSS Serviços de colaboração e implantação ou VantagensIBM SPSS Modeler.
- **Nós de Origem, de Processo e de Saída do IBM SPSS Modelador.** Descrições de todos os nós usados para ler, processar e emitir dados em diferentes formatos. Efetivamente, isso significa todos os nós além dos de modelagem.
- **Nós de Modelagem do IBM SPSS Modelador.** Descrições de todos os nós usados para criar modelos de mineração de dados. OIBM SPSS Modelador oferece uma variedade de métodos de modelagem tomados do aprendizado de máquina, inteligência artificial e estatística.
- **Guia de Aplicativos do IBM SPSS Modelador.** Os exemplos neste guia fornecem introduções sintetizadas e direcionadas para técnicas e métodos de modelagem específicos. Uma versão online deste guia também está disponível no menu Ajuda. Veja o tópico “Exemplos de Aplicação” na página 4 para obter mais informações.
- **Script e Automação Python do IBM SPSS Modelador.** Informações sobre como automatizar o sistema por meio de script Python, incluindo as propriedades que podem ser usadas para manipular nós e fluxos.
- **Guia de Implementação do IBM SPSS Modelador.** Informações sobre a execução de fluxos IBM SPSS Modelador como etapas de processamento de tarefas sob IBM SPSS Gerente de implantação.
- **Guia de Mineração Dentro do Banco de Dados do IBM SPSS Modelador.** Informações sobre como usar o poder do seu banco de dados para melhorar o desempenho e ampliar o intervalo de capacidades analíticas por meio de algoritmos de terceiros.
- **Guia de Desempenho e de Administração do ServidorIBM SPSS Modeler.** Informações sobre como configurar e administrar o ServidorIBM SPSS Modeler.

- **Guia do Usuário do IBM SPSS Gerente de implantação.** Informações sobre o uso da interface de usuário do console de administração incluídas no aplicativo Gerente de implantação para monitoramento e configuração ServidorIBM SPSS Modeler.
- **IBM SPSS Modelador Guia CRISP-DM.** Guia passo a passo para o uso da metodologia CRISP-DM para mineração de dados com ModeladorSPSS.
- **IBM SPSS Lote do modelador User's Guide.** Guia completo para o uso do IBM SPSS Modelador no modo em lote, incluindo detalhes da execução do modo em lote e argumentos de linha de comandos. Este guia está disponível somente em formato PDF.

SPSS Modeler Premium documentação

O conjunto de documentações do SPSS Modeler Premium (excluindo instruções de instalação) é o seguinte.

- **SPSS Modelador de análise de texto User's Guide.** Informações sobre o uso de analítica de texto com ModeladorSPSS, cobrindo os nós de mineração de texto, ambiente de trabalho interativo, modelos e outros recursos.

Exemplos de Aplicação

Enquanto as ferramentas de mineração de dados no ModeladorSPSS podem ajudar a resolver uma ampla variedade de negócios e problemas organizacionais, os exemplos de aplicativos fornecem introduções breves e destinadas aos métodos e técnicas de modelagem específicos. Os conjuntos de dados utilizados aqui são muito menores do que as enormes lojas de dados gerenciadas por alguns mineiros de dados, mas os conceitos e métodos que estão envolvidos são escaláveis para aplicações do mundo real.

Para acessar os exemplos, clique em **Exemplos de aplicativos** no menu Ajuda em ModeladorSPSS.

Os arquivos de dados e os fluxos de amostra são instalados na pasta `Demos` no diretório de instalação do produto. Para obter mais informações, consulte “Pasta Demos” na página 4.

Exemplos de modelagem da base de dados. Consulte os exemplos no *Guia de Mineração dentro do Banco de Dados do IBM SPSS Modelador*.

Exemplos de script. Consulte os exemplos no *Guia de Script e Automação do IBM SPSS Modelador*.

Pasta Demos

Os arquivos de dados e fluxos de amostra que são utilizados com os exemplos de aplicação são instalados na pasta `Demos` sob o diretório de instalação do produto (por exemplo: `C:\Program Files\IBM\SPSS\Modeler\<version>\Demos`). Esta pasta também pode ser acessada a partir do grupo de programas IBM ModeladorSPSS no menu Iniciar do Windows, ou clicando em `Demos` na lista de diretórios recentes na caixa de diálogo **Arquivo** > **Open Stream** .

Rastreamento de Licença

Quando você usa o ModeladorSPSS, o uso sob licença é controlado e registrado em intervalos regulares. As métricas de licença que são registradas são *AUTHORIZED_USER* e *CONCURRENT_USER* e o tipo de métrica que é registrado depende do tipo de licença que você possui para o ModeladorSPSS.

Os arquivos de log que são produzidos podem ser processados pelo IBM License Metric Tool, do qual é possível gerar relatórios de uso sob licença.

Os arquivos de log de licença são criados no mesmo diretório onde os arquivos de log do Client log do ModeladorSPSS são registrados (por padrão, `%ALLUSERSPROFILE%/IBM/SPSS/Modeler/<version>/log`).

Capítulo 2. Visão geral do produto

Introdução

Como um aplicativo de mineração de dados, o IBM SPSS Modelador oferece uma abordagem estratégica para localizar relacionamentos úteis em grandes conjuntos de dados. Em contraste com métodos estatísticos mais tradicionais, você não precisará necessariamente saber o que está procurando ao iniciar. Você pode explorar os dados, ajustando diferentes modelos e investigando diferentes relacionamentos até localizar informações úteis.

Iniciando o IBM SPSS Modelador

Para iniciar o aplicativo, clique em:

Iniciar > **[Todos] Programas** > **IBM SPSS Modeler < version>** > **IBM SPSS Modeler < version>**

A janela principal é exibida após alguns segundos.

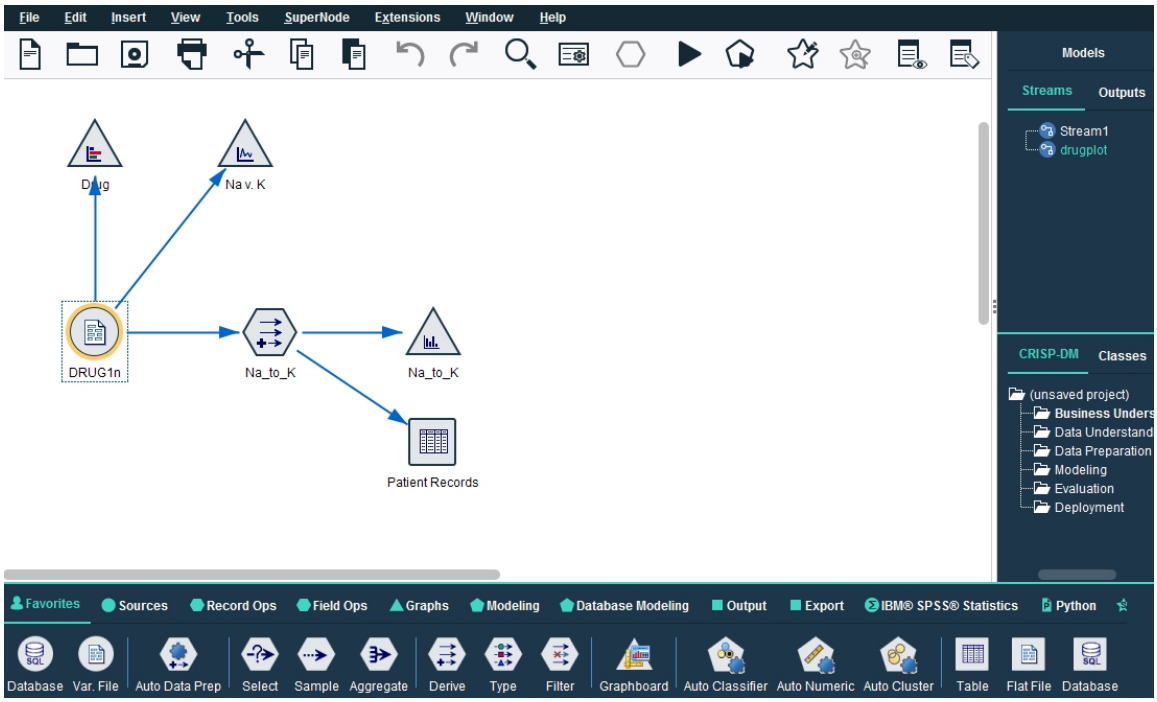


Figura 1. IBM SPSS Modelador janela principal do aplicativo

Ativando a partir da linha de comandos

Você pode usar a linha de comando do seu sistema operacional para iniciar o ' IBM SPSS Modelador da seguinte forma.

Microsoft Windows

1. Em um computador no qual o IBM SPSS Modelador está instalado, abra um DOS, um prompt de comandos ou uma janela.
2. Alterne para o caminho de instalação de IBM SPSS Modelador (por exemplo, `[Installpath]\Program Files\IBM\SPSS\Modeler\19.0\bin`).

3. Para iniciar a interface ' IBM SPSS Modelador no modo interativo, digite o comando ' `modelerclient` seguido dos argumentos necessários; por exemplo:

```
modelerclient -stream report.str -execute
```

Você pode usar os argumentos disponíveis (sinalizadores) para se conectar a um servidor, carregar fluxos, executar scripts ou especificar outros parâmetros, conforme necessário.

Mac OS

1. Localize o caminho do comando Mac OS para IBM SPSS Modelador (por exemplo, `[Installpath]/Applications/IBM/SPSS/Modeler/19.0/IBM SPSS Modeler.app/Contents/MacOS`).
2. Para iniciar a interface ' IBM SPSS Modelador no modo interativo, execute o comando modeler seguido dos argumentos necessários; por exemplo:

```
./modeler -stream report.str -execute
```

Conectando-se ao ServidorIBM SPSS Modeler

IBM SPSS Modelador pode ser executado como um aplicativo independente, ou como um cliente conectado diretamente ao ServidorIBM SPSS Modeler, ou a um ServidorIBM SPSS Modeler ou cluster de servidores por meio do plug-in Coordinator of Processes do IBM SPSS Serviços de colaboração e implantação. O status da conexão atual é exibido na parte inferior esquerda da janela do IBM SPSS Modelador.

Sempre que você desejar se conectar a um servidor, é possível inserir manualmente o nome do servidor ao qual deseja se conectar ou selecionar um nome que você definiu anteriormente. No entanto, se você tiver o IBM SPSS Serviços de colaboração e implantação, é possível procurar em uma lista de servidores ou clusters de servidores na caixa de diálogo Login do Servidor. A capacidade de navegar pelos serviços do Estatísticas em execução em uma rede é disponibilizada por meio do Coordinator of Processes.

Para se conectar a um servidor

1. No menu Ferramentas, clique em **Login do Servidor**. A caixa de diálogo Login do Servidor se abre. Alternativamente, dê um clique duplo na área de status da conexão da janela IBM SPSS Modelador.
2. Usando a caixa de diálogo, especifique opções para se conectar ao computador servidor local ou selecione uma conexão da tabela.
   - Clique em **Incluir** ou **Editar** para incluir ou editar uma conexão. Consulte o tópico “Como adicionar e Editar a Conexão ServidorIBM SPSS Modeler” na página 7 para obter mais informações.
   - Clique em **Procurar** para acessar um servidor ou cluster de servidores no Coordinator of Processes. Veja o tópico “Procurando por Servidores em IBM SPSS Serviços de colaboração e implantação” na página 7 para obter mais informações.

   Tabela de servidores. Essa tabela contém o conjunto de conexões de servidor definidas. A tabela exibe a conexão padrão, o nome do servidor, a descrição e o número da porta. É possível incluir manualmente uma nova conexão, bem como selecionar ou procurar uma conexão existente. Para configurar um determinado servidor como conexão padrão, marque a caixa de seleção na coluna Padrão na tabela para a conexão.

   Caminho de dados padrão. Especifique um caminho usado para dados no computador servidor. Clique no botão de reticências **(...)** para navegar para o local necessário.

   Configurar credenciais. Deixe essa caixa desmarcada para ativar a variável **conexão única**, que tenta efetuar seu login no servidor usando detalhes de nome de usuário e senha do computador local. Se uma conexão única não for possível ou se você marcar essa caixa para desativar a conexão única (por exemplo, para efetuar login em uma conta do administrador), os campos a seguir serão ativados para você inserir suas credenciais.

   ID do Usuário. Insira o nome de usuário com o qual efetuar logon no servidor.

Senha. Insira a senha associada ao nome de usuário especificado.

Domínio. Especifique o domínio usado para efetuar logon no servidor. Um nome de domínio só é necessário quando o computador servidor está em um domínio do Windows diferente daquele do computador cliente.

3. Clique em **OK** para concluir a conexão.

Para se desconectar de um servidor

1. No menu Ferramentas, clique em **Login do Servidor**. A caixa de diálogo Login do Servidor se abre. Alternativamente, dê um clique duplo na área de status da conexão da janela IBM SPSS Modelador.
2. Na caixa de diálogo, selecione o Servidor Local e clique em **OK**.

Como adicionar e Editar a Conexão ServidorIBM SPSS Modeler

É possível editar ou incluir manualmente uma conexão do servidor na caixa de diálogo Login do Servidor. Clicando em Incluir, é possível acessar uma caixa de diálogo Incluir/Editar Servidor vazia na qual é possível inserir detalhes da conexão do servidor. Com a seleção de uma conexão existente e um clique em Editar na caixa de diálogo Login do Servidor, a caixa de diálogo Incluir/Editar é aberta com os detalhes para essa conexão, de modo que seja possível fazer quaisquer mudanças.

Nota: Não é possível editar uma conexão do servidor que foi incluída do IBM SPSS Serviços de colaboração e implantação, já que o nome, a porta e outros detalhes são definidos no IBM SPSS Serviços de colaboração e implantação. A melhor prática determina que as mesmas portas devem ser usadas para comunicação com IBM SPSS Serviços de colaboração e implantação e ModeladorSPSS Client. Isso pode ser configurado como `max_server_port` e `min_server_port` no arquivo options.cfg.

Para incluir conexões do servidor

1. No menu Ferramentas, clique em **Login do Servidor**. A caixa de diálogo Login do Servidor se abre.
2. Nesta caixa de diálogo, clique em **Incluir**. A caixa de diálogo Incluir/Editar Servidor de Login do Servidor é aberta.
3. Insira os detalhes de conexão do servidor e clique em **OK** para salvar a conexão e retornar para a caixa de diálogo Login do Servidor.

- **Servidor.** Especifique um servidor disponível ou selecione um na lista. O computador servidor pode ser identificado por um nome alfanumérico (por exemplo, *myserver*) ou um endereço IP designado ao computador servidor (por exemplo, 202.123.456.78).
- **Porta.** Forneça o número da porta no qual o servidor está atendendo. Se o padrão não funcionar, peça ao administrador do sistema o número da porta correto.
- **Descrição.** Insira uma descrição opcional para essa conexão do servidor.
- **Assegurar conexão segura (usar SSL).** Especifica se uma conexão SSL (**Secure Sockets Layer**) deve ser usada. O SSL é um protocolo usado comumente para proteger os dados enviados em uma rede. Para usar essa variável, o SSL deve ser ativado no servidor hospedando o ServidorIBM SPSS Modeler. Se necessário, entre em contato com o administrador local para obter detalhes.

Para editar conexões do servidor

1. No menu Ferramentas, clique em **Login do Servidor**. A caixa de diálogo Login do Servidor se abre.
2. Nessa caixa de diálogo, selecione a conexão que deseja editar e clique em **Editar**. A caixa de diálogo Incluir/Editar Servidor de Login do Servidor é aberta.
3. Altere os detalhes de conexão do servidor e clique em **OK** para salvar as mudanças e retornar para a caixa de diálogo Login do Servidor.

Procurando por Servidores em IBM SPSS Serviços de colaboração e implantação

Em vez de inserir uma conexão do servidor manualmente, é possível selecionar um servidor ou cluster de servidores disponíveis na rede por meio do Coordinator of Processes, disponível no IBM SPSS Serviços

de colaboração e implantação. Um cluster de servidores é um grupo de servidores a partir do qual o Coordinator of Processes determina o servidor mais adequado para responder a uma solicitação de processamento.

Embora seja possível incluir manualmente servidores na caixa de diálogo Login do Servidor, procurar por servidores disponíveis permite conectar-se aos servidores sem precisar saber o nome e o número da porta corretos do servidor. Estas informações são fornecidas automaticamente. Entretanto, você ainda precisará das informações de logon corretas, como nome de usuário, domínio e senha.

Nota: Se você não tiver acesso à capacidade Coordinator of Processes, ainda será possível inserir manualmente o nome do servidor ao qual deseja se conectar ou selecionar um nome que você definiu anteriormente. Veja o tópico "Como adicionar e Editar a Conexão ServidorIBM SPSS Modeler" na página 7 para obter mais informações.

Para procurar servidores e clusters

1. No menu Ferramentas, clique em **Login do Servidor**. A caixa de diálogo Login do Servidor se abre.
2. Nesta caixa de diálogo, clique em **Procurar** para abrir a caixa de diálogo Procurar por Servidores. Se você não tiver efetuado logon no IBM SPSS Serviços de colaboração e implantação quando tentou navegar no Coordinator of Processes, será solicitado que faça isso agora.
3. Selecione o servidor ou cluster de servidores da lista.
4. Clique em **OK** para fechar a caixa de diálogo e incluir essa conexão na tabela na caixa de diálogo Login do Servidor.

Conectando-se ao Servidor analítico

Se você tiver vários Servidor analíticos disponíveis, você pode usar o diálogo Conexão do Servidor Analítico para definir mais de um servidor para uso em IBM SPSS Modelador. O seu administrador pode já ter configurado um padrão Servidor analítico no arquivo `<Modeler_install_path>/config/options.cfg` . Mas você também pode usar outros servidores disponíveis após defini-los. Por exemplo, ao usar os nós Servidor analítico Source e Export, você pode querer usar diferentes Servidor analítico conexões em diferentes ramos de um fluxo para que quando cada ramificação execute ele use seus próprios Servidor analítico e nenhum dado será puxado para o ServidorIBM SPSS Modeler. Note que se uma ramificação contém mais de uma conexão Servidor analítico , os dados serão retirados dos Servidor analíticos para o ServidorIBM SPSS Modeler.

Para criar uma nova conexão Servidor analítico , vá em **Ferramentas** > **Conexões do Servidor Analítico** e forneça as informações necessárias nas seguintes seções do diálogo.

Conexão

URL. Digite URL para o Servidor analítico no formato `https://hostname:port/contextroot`, em que `hostname` é o endereço IP ou o nome do host do Servidor analítico, `port` é o número da porta e `contextroot` é a raiz do contexto do Servidor analítico.

Inquilino. Digite o nome do locatário que o ServidorIBM SPSS Modeler é um membro. Entre em contato com o administrador se você não conhece o inquilino.

Autenticação

Modo. Selecione a partir dos seguintes modos de autenticação.

- **Nome de usuário e senha** requer que você digite o nome de usuário e a senha.
- **Credencial Armazenada** requer que você selecione uma credencial a partir do IBM SPSS Repositório de serviços de colaboração e implantação.
- **Kerberos** requer que você digite o nome principal do serviço e o caminho de arquivo config. Entre em contato com o administrador se você não souber essas informações.

Nome de usuário. Digite o nome de usuário Servidor analítico .

Resmas. Selecione o reino a ser usado para a conexão Servidor analítico .

Senha. Digite a senha do Servidor analítico .

Conectar. Clique em **Conectar** para testar a nova conexão.

Conexões

Depois de especificar as informações acima e clicar em **Connect**, a conexão será adicionada a esta tabela Conexões. Se você precisar remover uma conexão, selecione-a e clique em **Remover**.

Se o seu administrador definiu uma conexão padrão do Servidor analítico no arquivo `options.cfg` , você pode clicar em **Adicionar conexão padrão** para adicioná-lo às suas conexões disponíveis também. Você será solicitado o nome de usuário e a senha.

Alterando o diretório temporário

Algumas operações que são feitas pelo ServidorIBM SPSS Modeler podem requerer que arquivos temporários sejam criados Por padrão, o IBM SPSS Modelador usa o diretório temporário do sistema para criar arquivos temp. É possível alterar a localização do diretório temporário usando os passos a seguir.

1. Crie um novo diretório chamado `spss` e um subdiretório chamado `servertemp`.
2. Edite `options.cfg`, localizado no diretório `/config` do diretório de instalação do IBM SPSS Modelador. Edite o parâmetro `temp_directory` neste arquivo para ler: `temp_directory, "C:/spss/servertemp"`.
3. Reinicie o serviço ServidorIBM SPSS Modeler. Isso pode ser feito clicando na guia **Serviços** do Painel de Controle do Windows. Pare o serviço e, em seguida, inicie-o para ativar as mudanças feitas. Reiniciar a máquina também reinicia o serviço.

Todos os arquivos temp agora são gravados neste novo diretório

Nota: Barras devem ser usadas.

Diretório temporário para visualização de dados

Para o serviço de visualização de dados, conclua as etapas a seguir para configurar o diretório temporário

1. Crie o diretório temp `D:/SPSSTemp`
2. Os arquivos temp da visualização de dados são controlados por `-Djava.io.tmpdir=D:/SPSSTemp` Inclua essa opção no script de início para visualização de dados no seguinte arquivo: `${ModelerInstallDir}/dataview/start_graph_micro_service.sh`.
3. Inicie o IBM SPSS Modelador.

Arquivos agora são gravados em `D:/SPSSTemp`.

Início Múltiplo IBM SPSS Modelador Sessões

Se você precisar ativar mais de uma sessão do IBM SPSS Modelador por vez, deve-se fazer algumas mudanças nas configurações do IBM SPSS Modelador e do Windows. Por exemplo, talvez seja necessário fazer isso se você tiver duas licenças de servidor separadas e desejar executar dois fluxos com relação a dois servidores diferentes a partir do mesmo computador cliente.

Para ativar várias sessões do IBM SPSS Modelador:

1. Clique em:

   Iniciar > **[Todos] Programas** > **IBM SPSS Modeler**
2. No atalho IBM SPSS Modelador (aquele com o ícone), clique com o botão direito do mouse e selecione **Propriedades**.
3. Na caixa de texto **Resposta**, inclua `-noshare` no final da sequência de caracteres.

IBM SPSS Modelador Interface em uma Glance

Em cada ponto no processo de mineração de dados, a interface do IBM SPSS Modelador fácil de usar convida seus conhecimentos de negócios específicos. Os algoritmos de modelagem, como predição, classificação, segmentação e detecção de associação, asseguram modelos poderosos e exatos. Os resultados do modelo podem ser facilmente implementados e lidos nos bancos de dados, IBM SPSS Estatísticas e em uma grande variedade de outros aplicativos.

O trabalho com o IBM SPSS Modelador é um processo de três passos de trabalho com dados.

- Primeiro, você lê dados no IBM SPSS Modelador.
- Depois, você executa os dados por meio de uma série de manipulações.
- Por fim, você envia os dados para um destino.

Essa sequência de operações é conhecida como **fluxo de dados**, pois os dados fluem registro por registro, desde a origem, passando por cada manipulação e, por fim, até seu destino--um modelo ou tipo de saída de dados.

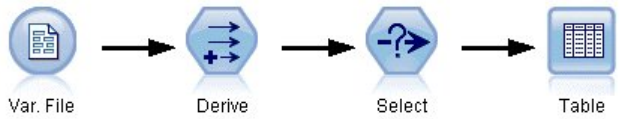


Figura 2. Um fluxo simples

IBM SPSS Modelador tela de fluxo

A tela de fluxo é a maior área da janela do IBM SPSS Modelador e é onde você construirá e manipulará fluxos de dados.

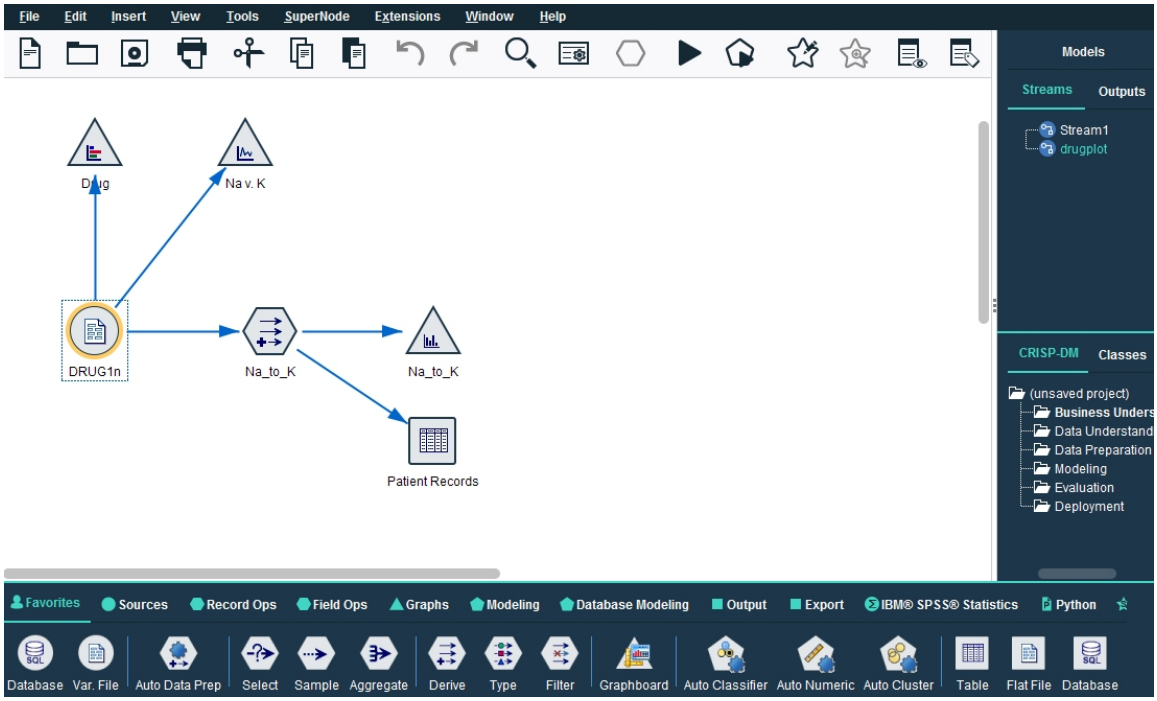


Figura 3. Área de trabalho do IBM SPSS Modelador (visualização padrão)

Fluxos são criados desenhando diagramas de operações de dados relevantes para seus negócios na tela principal da interface. Cada operação é representada por um ícone ou **nó**, e os nós são vinculados em um **fluxo** representando o fluxo de dados em cada operação.

É possível trabalhar com vários fluxos de uma vez no IBM SPSS Modelador, ou na mesma tela de fluxo ou abrindo uma nova. Durante uma sessão, fluxos são armazenados no gerenciador de Fluxos, no lado superior direito da janela do IBM SPSS Modelador.

Nota: Se estiver usando um MacBook com a configuração **Force Click and haptic feedback** do trackpad integrado ativada, arrastar e soltar nós da paleta de nós para a tela de fluxo pode resultar na inclusão de nós duplicados na tela. Para evitar essa questão, recomendamos a desativação da preferência do sistema **Force Click and haptic feedback** trackpad.

Paleta de nós

A maioria das ferramentas de dados e modelagem em ModeladorSPSS estão disponíveis a partir da *Paleta de Nodes*, através da parte inferior da janela abaixo da tela do fluxo.

Por exemplo, a guia paleta **Registrar Ops** contém nós que você pode usar para executar operações nos dados *registros*, como selecionar, mesclar e anexar.

Para adicionar nós à tela, dê um duplo clique em ícones da Paleta de Nós ou arraste-os para a tela. Depois conecte-os para criar um *fluxo*, representando o fluxo de dados.

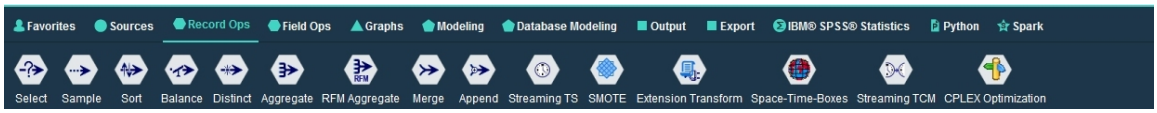


Figura 4. Guia Operações de Registro na paleta de nós

Cada guia da paleta contém uma coleção de nós relacionados usados para diferentes fases de operações de fluxo, como:

- **Fontes** nós trazem dados em ModeladorSPSS.
- **Record Ops** nós executam operações em dados *records*, como selecionar, mesclar e anexar.
- **Ops de campo** nós executam operações em dados *campos*, como filtragem, derivando novos campos e determinando o nível de medição para determinados campos.
- **Gráficos** nós exibem graficamente os dados antes e depois da modelagem. Gráficos incluem plots, histogramas, nós da web e gráficos de avaliação.
- **Modelagem** nós utilizamos os algoritmos de modelagem disponíveis em ModeladorSPSS, tais como redes neurais, árvores de decisão, algoritmos de clustering e sequenciamento de dados.
- **Modelagem de Banco de Dados** nós usamos os algoritmos de modelagem disponíveis em Microsoft SQL Server, IBM Db2, e bancos de dados Oracle e Netezza .
- **Nós de saída** produzem várias saídas para dados, gráficos e resultados de modelos que podem ser visualizados em ModeladorSPSS.
- **Os nós de exportação** produzem várias saídas que podem ser visualizadas em aplicativos externos, como IBM SPSS Coleta de dados ou Excel.
- **IBM SPSS Estatísticas** nós importar dados de, ou exportar dados para, IBM SPSS Estatísticas , bem como executar procedimentos IBM SPSS Estatísticas .
- Os nós **Python** podem ser usados para executar algoritmos Python .
- Os nós **Spark** podem ser usados para executar algoritmos de Spark.

Conforme você se familiariza com ModeladorSPSS, é possível customizar o conteúdo da paleta para seu próprio uso.

No lado esquerdo da Paleta de Nós, é possível filtrar os nós que são exibidos selecionando Supervised, Association ou Segmentação.

Localizada abaixo da Paleta de Nós, uma área de janela de relatório fornece feedback sobre o progresso de várias operações, como quando os dados estão sendo lidos no fluxo de dados. Também localizada abaixo da Paleta de Nós, uma área de janela de status fornece informações sobre o que o aplicativo está fazendo no momento, bem como indicações de quando um feedback do usuário é necessário.

Nota: Se estiver usando um MacBook com a configuração **Force Click and haptic feedback** do trackpad integrado ativada, arrastar e soltar nós da paleta de nós para a tela de fluxo pode resultar na inclusão de nós duplicados na tela. Para evitar essa questão, recomendamos a desativação da preferência do sistema **Force Click and haptic feedback** trackpad.

IBM SPSS Modelador Gerenciadores

Na parte superior direita da janela está a área de janela de gerenciadores. Ela possui três guias, que são usadas para gerenciar fluxos, saída e modelos.

É possível usar a guia Fluxos para abrir, renomear, salvar e excluir os fluxos criados em uma sessão.

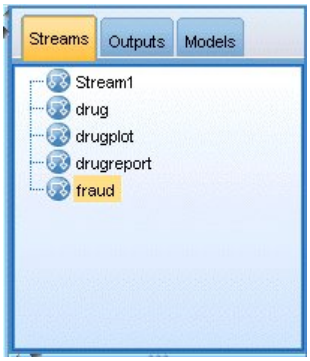


Figura 5. Guia Fluxos

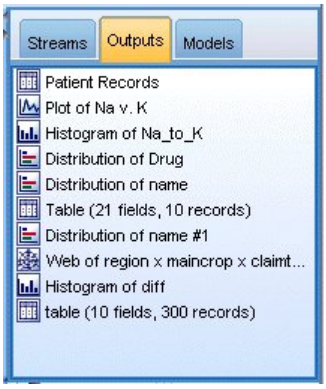


Figura 6. Guia Saídas

A guia Saídas contém uma variedade de arquivos, como gráficos e tabelas, produzidos por operações de fluxo no IBM SPSS Modelador. É possível exibir, salvar, renomear e fechar as tabelas, gráficos e relatórios listados nessa guia.

Figura 7. Guia Modelos contendo nuggets de modelo

A guia Modelos é a mais poderosa das guias de gerenciadores. Essa guia contém todos os **nuggets** de modelo, que contêm os modelos gerados no IBM SPSS Modelador, para a sessão atual. Esses modelos podem ser navegados diretamente a partir da guia Modelos ou incluídos no fluxo na tela.

IBM SPSS ModeladorProjetos

No lado inferior direito da janela está a área de janela do projeto, usada para criar e gerenciar **projetos** de mineração de dados (grupos de arquivos relacionados a uma tarefa de mineração de dados). Há duas maneiras de se visualizar projetos que você cria no IBM SPSS Modelador—na visualização Classes e na visualização CRISP-DM.

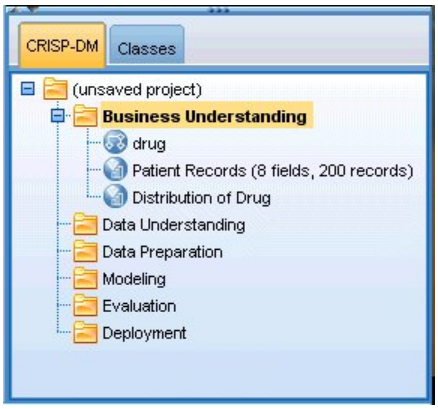


Figura 8. Visualização CRISP-DM

A guia CRISP-DM fornece uma maneira de organizar projetos de acordo com o processo padrão de vários segmentos de mercados para mineração de dados, uma metodologia não proprietária comprovada pela indústria. Tanto para mineradores de dados iniciantes quanto para os experientes, o uso da ferramenta CRISP-DM ajudará você a organizar e comunicar melhor seus esforços.

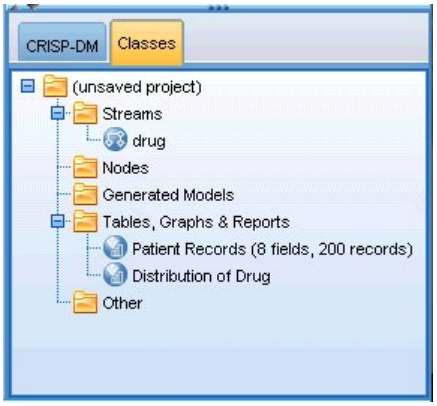


Figura 9. Visualização Classes

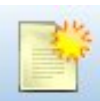
A guia Classes fornece uma maneira de organizar seu trabalho no IBM SPSS Modelador categoricamente- por tipos de objetos criados. Essa visualização é útil durante a realização de inventário de dados, fluxos e modelos.

IBM SPSS Modelador Barra de ferramentas

Na parte superior da janela do IBM SPSS Modelador, você localizará uma barra de ferramentas de ícones que fornecem inúmeras funções úteis. A seguir estão os botões da barra de ferramentas e suas funções.

Criar novo fluxo

Abrir fluxo

	Salvar fluxo		Imprimir fluxo atual
	Cortar e mover para a área de transferência		Copiar para área de transferência
	Colar seleção		Desfazer última ação
	Refazer		Procurar nós
	Edite propriedades do fluxo		Visualizar geração de SQL
	Executar fluxo atual		Executar seleção de fluxo
	Parar fluxo (Ativo somente enquanto o fluxo está em execução)		Incluir Supernó
	Aumentar zoom (Somente SuperNodes)		Diminuir zoom (Somente SuperNodes)
	Nenhuma marcação no fluxo		Inserir comentário
	Ocultar marcação de fluxo (se houver)		Mostrar marcação de fluxo oculta
	Abrir fluxo em VantagensIBM SPSS Modeler		

Marcação de fluxo consiste em comentários do fluxo, ligações de modelo e indicações de ramificação de escoragem.

Ligações de modelo são descritas no guia *Nós de Modelagem do IBM SPSS*.

Customizando a barra de ferramentas

É possível alterar vários aspectos da barra de ferramentas, como:

- Se ela é exibida
- Se os ícones têm dicas de ferramentas disponíveis
- Se ela usa ícones grandes ou pequenos

Para ativar e desativar a exibição da barra de ferramentas:

1. No menu principal, clique em:

   Visualizar > **Barra de Ferramentas** > **Display**

Para alterar as configurações de dica de ferramenta ou tamanho do ícone:

1. No menu principal, clique em:

   View > **Barra de Ferramentas** > **customizar**

Clique em **Mostrar Dicas de Ferramenta** ou **Botões Grandes** conforme necessário.

Como personalizar a janela IBM SPSS Modelador

Usando as divisórias entre várias partes da interface do ModeladorSPSS, é possível redimensionar ou fechar ferramentas para atender às suas preferências. Por exemplo, se você estiver trabalhando com um grande fluxo, é possível usar as pequenas setas localizadas em cada divisória para fechar a paleta do nó, a área de janela de gerenciadores e a área de janela de projeto. Isso maximiza a tela de fluxo, fornecendo um espaço de trabalho suficiente para fluxos grandes ou para vários fluxos.

Alternativamente, no menu Visualizar, clique em **Paleta de Nós**, **Gerenciadores** ou **Projeto** para ativar ou desativar a exibição desses itens.

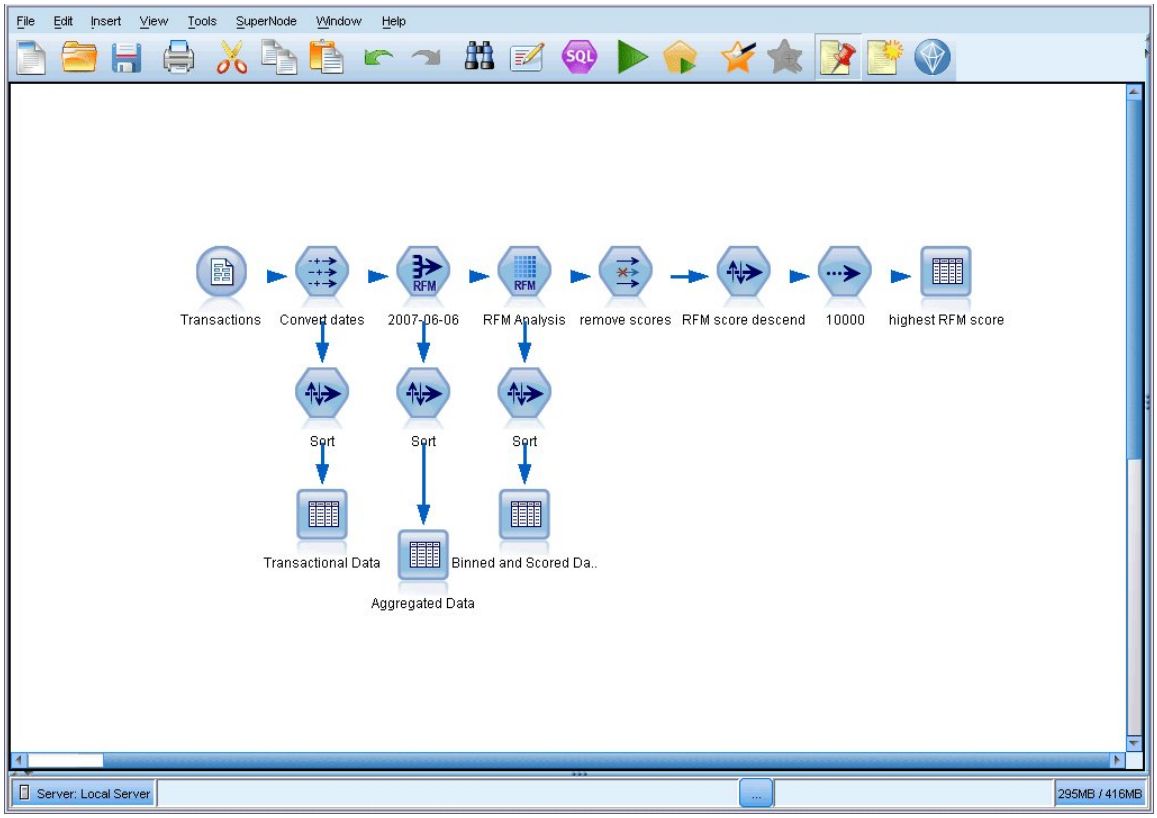


Figura 10. Tela de fluxo maximizada

Como uma alternativa para o fechamento da paleta de nós e para as áreas de janela de gerenciador e projeto, é possível usar a tela de fluxo como uma página rolável se movendo vertical e horizontalmente com as barras de rolagem ao lado e na parte inferior da janela do ModeladorSPSS.

Também é possível controlar a exibição da marcação da tela, o que consiste em comentários do fluxo, ligações de modelo e indicações de ramificação de escoragem. Para ativar ou desativar essa exibição, clique em:

Exibir > **marcação de fluxo**

Alterando o tamanho do ícone para um fluxo

É possível alterar o tamanho dos ícones de fluxo das seguintes formas.

- Por meio de uma configuração de propriedades de fluxo
- Por meio de um menu pop-up no fluxo
- Usando o teclado

É possível escalar a visualização de fluxo inteira para um de vários tamanhos entre 8% e 200% do tamanho do ícone padrão.

Para escalar o fluxo inteiro (método propriedades do fluxo)

1. No menu principal, escolha

   Ferramentas > **Propriedades do Fluxo** > **Opções** > **Layout.**
2. Escolha o tamanho desejado no menu Tamanho do Ícone.
3. Clique em **Aplicar** para ver o resultado.
4. Clique em **OK** para salvar a mudança.

Para escalar o fluxo inteiro (método menu)

1. Clique com o botão direito no segundo plano do fluxo na tela.
2. Escolha **Tamanho do Ícone** e selecione o tamanho desejado.

Para escalar o fluxo inteiro (método teclado)

1. Pressione Ctrl + [-] no teclado principal para diminuir zoom para o próximo menor tamanho.
2. Pressione Ctrl + Shift + [+] no teclado principal para aumentar zoom para o próximo maior tamanho.

   Observe que este método de zoom em pode não funcionar dependendo do seu sistema operacional e teclado utilizado.

Essa variável é particularmente útil para se ter uma visualização geral de um fluxo complexo. Também é possível usá-la para minimizar o número de páginas necessárias para se imprimir um fluxo.

Usando o Mouse em IBM SPSS Modelador

Os usos mais comuns do mouse no IBM SPSS Modelador incluem o seguinte:

- **Clique único.** Use o botão direito ou esquerdo do mouse para selecionar opções dos menus, abrir menus pop-up e acessar vários outros controles e opções padrão. Clique e segure o botão para mover e arrastar nós.
- **Clique duplo.** Dê um clique duplo usando o botão esquerdo do mouse para colocar nós na tela de fluxo e editar nós existentes.
- **Clique do meio.** Clique com o botão do meio do mouse e arraste o cursor para conectar nós na tela de fluxo. Dê um clique duplo no botão do meio do mouse para desconectar um nó. Se você não tiver um mouse com três botões, é possível simular essa variável pressionando a tecla Alt ao clicar e arrastar o mouse.

Usando teclas de atalho

Muitas operações de programação visual no IBM SPSS Modelador têm teclas de atalho associadas a elas. Por exemplo, é possível excluir um nó clicando nele e pressionando a tecla Delete em seu teclado. Da mesma forma, é possível salvar rapidamente um fluxo pressionando a tecla S enquanto você mantém pressionada a tecla Ctrl. Comandos de controle como esse são indicados por uma combinação de Ctrl e outra tecla -- por exemplo, Ctrl+S.

Há inúmeras teclas de atalho usadas em operações padrão do Windows, como Ctrl+X para cortar. Esses atalhos são suportados no IBM SPSS Modelador com os seguintes atalhos específicos do aplicativo.

Nota: Em alguns casos, antigas teclas de atalho usadas em IBM SPSS Modelador entram em conflito com teclas de atalho padrão do Windows. Esses atalhos antigos são suportados com a adição da tecla Alt. Por exemplo, Ctrl+Alt+C pode ser usado para ativar e desativar o cache.

Tabela 1. Teclas de atalho suportadas

Tecla de Atalho	**Função**
Ctrl+A	Selecionar todos
Ctrl+X	Cortar
Ctrl+N	Novo fluxo
Ctrl+O	Abrir fluxo
Ctrl+P	Imprimir
Ctrl+C	Copiar
Ctrl+V	Colar
Ctrl+Z	Desfazer
Ctrl+Q	Selecionar todos os nós de recebimentos de dados do nó selecionado
Ctrl+W	Cancelar todos os nós de recebimento de dados (alterna-se com Ctrl+Q)
Ctrl+E	Executar a partir do nó selecionado
Ctrl+S	Salvar fluxo atual
Alt+Teclas de Seta	Move os nós selecionados na tela de fluxo na direção da tecla usada
Shift+F10	Abre o menu pop-up para o nó selecionado

Tabela 2. Atalhos suportados para antigas teclas de atalho

Tecla de Atalho	**Função**
Ctrl+Alt+D	Duplicar nó
Ctrl+Alt+L	Carregar nó
Ctrl+Alt+R	Renomear nó
Ctrl+Alt+U	Criar nó de Entrada do Usuário
Ctrl+Alt+C	Ativar/desativar cache
Ctrl+Alt+F	Limpar o cache
Ctrl+Alt+X	Expandir Supernó
Ctrl+Alt+Z	Aumentar zoom/diminuir aumentar zoom
Excluir	Excluir nó ou conexão

Imprimindo

Os objetos a seguir podem ser impressos no IBM SPSS Modelador:

- Diagramas de fluxo
- Gráficos

- Tabelas
- Relatórios (no nó Relatório e Relatórios do Projeto)
- Scripts (das caixas de diálogo de propriedades do fluxo, Script Independente ou Script SuperNode)
- Modelos (navegadores Modelo, guias de caixa de diálogo com foco atual, visualizadores de árvore)
- Anotações (usando a guia Anotações para saída)

Para imprimir um objeto:

- Para imprimir sem visualizar, clique no botão Imprimir na barra de ferramentas.
- Para configurar a página antes da impressão, selecione **Configuração de Página** no menu Arquivo.
- Para visualizar antes de imprimir, selecione **Visualização de Impressão** no menu Arquivo.
- Para visualizar a caixa de diálogo de impressão padrão com opções para selecionar impressoras e especificar opções de aparência, selecione **Imprimir** no menu Arquivo.

automatizandoIBM SPSS Modelador

Como a mineração de dados avançada pode ser um processo complexo e, às vezes, longo, o IBM SPSS Modelador inclui vários tipos de codificação e o suporte de automação.

- **Control Language for Expression Manipulation** (CLEM) é uma linguagem para analisar e manipular os dados que fluem ao longo dos fluxos do IBM SPSS Modelador. Os mineradores de dados usam CLEM extensivamente em operações de fluxo para executar tarefas tão simples quanto derivar lucros dos dados de custo e renda ou tão complexas quanto transformar dados de log da web em um conjunto de campos e registros com informações utilizáveis.
- **Script** é uma ferramenta poderosa para automatizar processos na interface com o usuário. Scripts podem executar os mesmos tipos de ações que os usuários executam com um mouse ou teclado. Também é possível especificar saída e manipular modelos gerados.

Capítulo 3. Introdução à Modelagem

Um modelo é um conjunto de regras, fórmulas ou equações que podem ser utilizadas para prever um resultado com base em um conjunto de campos de entrada ou variáveis. Por exemplo, uma instituição financeira pode utilizar um modelo para prever se os solicitantes de empréstimo poderão representar um bom ou mau risco, com base nas informações que já se conhece sobre os solicitantes passados.

A capacidade de prever um resultado é o objetivo central de análise preditiva e entender o processo de modelagem é a chave para utilizar o IBM SPSS Modelador.

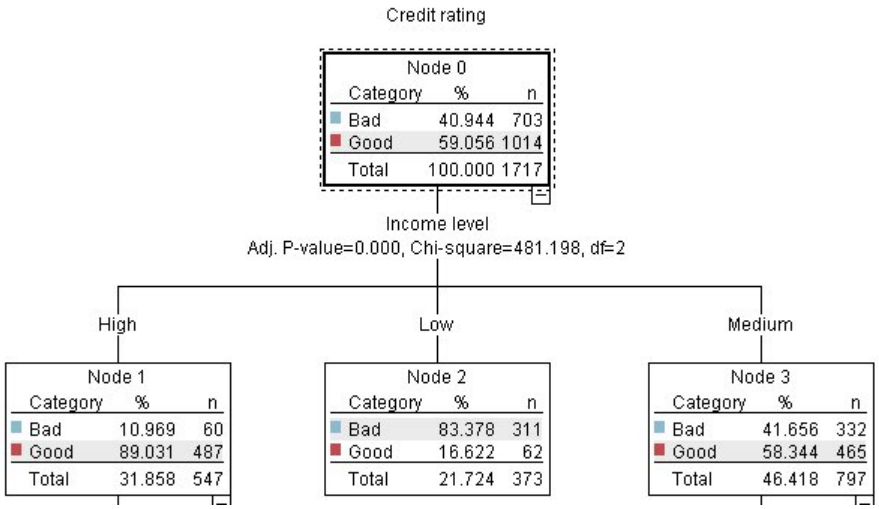


Figura 11. Um modelo de árvore de decisão simples

Esse exemplo utiliza um modelo de **árvore de decisão**, que classifica registros (e prediz uma resposta) usando uma série de regras de decisão, por exemplo:

```
IF income = Medium
AND cards <5
THEN -> 'Good'
```

Embora este exemplo use um modelo CHAID (Detecção de Interação Automática de Qui-quadrado), destina-se a ser uma introdução geral e a maioria dos conceitos se aplica amplamente a outros tipos de modelagem no IBM SPSS Modelador.

Para entender qualquer modelo, primeiro deve-se entender os dados que entram nele. Os dados neste exemplo contêm informações sobre os clientes de um banco. Os campos a seguir são utilizados:

Nome do campo	**Descrição**
Credit_rating	Classificação de crédito: 0=Bad, 1=Good, 9=missing values
Idade	Idade em anos
Receita	Nível de renda: 1=Low, 2=Medium, 3=High
Credit_cards	Número de cartões de crédito que possui: 1=Less than five, 2=Five or more
Educação	Nível de educação: 1=High school, 2=College
Car_loans	Número de empréstimos para compra de carro contraídos: 1=None or one, 2=More than two

O banco mantém um banco de dados de informações históricas sobre os clientes que contraíram empréstimos do banco, incluindo se eles pagaram os empréstimos (Classificação de crédito = Bom) ou

se ficaram inadimplentes (Classificação de crédito = Ruim). Utilizando esses dados existentes, o banco constrói um modelo que permitirá prever quão provavelmente futuros solicitantes de empréstimo se tornarão inadimplentes.

Utilizando um modelo de árvore de decisão, é possível analisar as características dos dois grupos de clientes e prever a probabilidade de inadimplência no empréstimo.

Esse exemplo usa o fluxo denominado *modelingintro.str*, disponível na pasta *Demos* sob a subpasta *streams*. O arquivo de dados é *tree_credit.sav*. Veja o tópico "Pasta Demos" na página 4 para obter mais informações.

Vamos dar uma olhada no fluxo.

1. Escolha o seguinte no menu principal:

   Arquivo > **Open Stream**
2. Clique no ícone de pepita de ouro na barra de ferramentas da caixa de diálogo Abrir e escolha a pasta Demos.
3. Clique duas vezes na pasta *streams*.
4. Clique duas vezes no arquivo denominado *modelingintro.str*.

Construindo o Fluxo

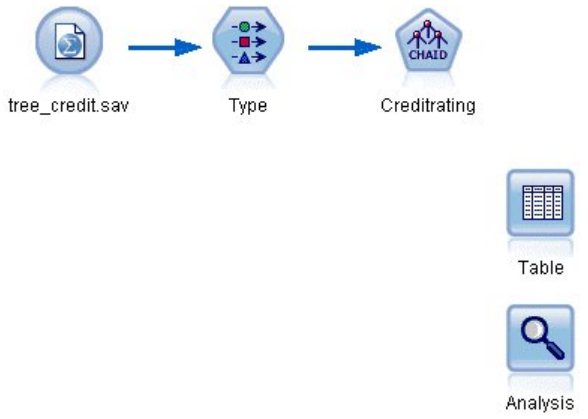


Figura 12. Fluxo de Modelagem

Para construir um fluxo que criará um modelo, pelo menos três elementos são necessários:

- Um nó de origem que lê dados a partir de alguma origem externa, nesse caso, um arquivo de dados do IBM SPSS Estatísticas .
- Uma origem ou nó Tipo que especifica as propriedades do campo, como nível de medição (o tipo de dados que o campo contém) e o papel de cada campo como um destino ou entrada na modelagem.
- Um nó de modelagem que gera um nugget do modelo quando o fluxo é executado.

Neste exemplo, estamos utilizando um nó de modelagem CHAID. O CHAID, ou Chi-squared Automatic Interaction Detection, é um método de classificação que constrói as árvores de decisão usando um tipo específico de estatísticas conhecido como estatísticas qui-quadrado para descobrir os melhores locais para fazer as divisões na árvore de decisão.

Se os níveis de medição forem especificados no nó de origem, o nó Tipo separado poderá ser eliminado. Funcionalmente, o resultado é o mesmo.

Este fluxo também tem os nós Tabela e Análise que serão usados para visualizar os resultados da escoragem após o nugget do modelo ter sido criado e incluído no fluxo.

O nó de origem Arquivo de Estatísticas lê dados no formato IBM SPSS Estatísticas a partir do arquivo de dados *tree_credit.sav*, que é instalado na pasta *Demos*. (Uma variável especial denominada

$CLEO_DEMOS é usada para referenciar essa pasta na instalação atual do IBM SPSS Modelador. Isso assegura que o caminho seja válido, independentemente da pasta de instalação ou da versão atual).

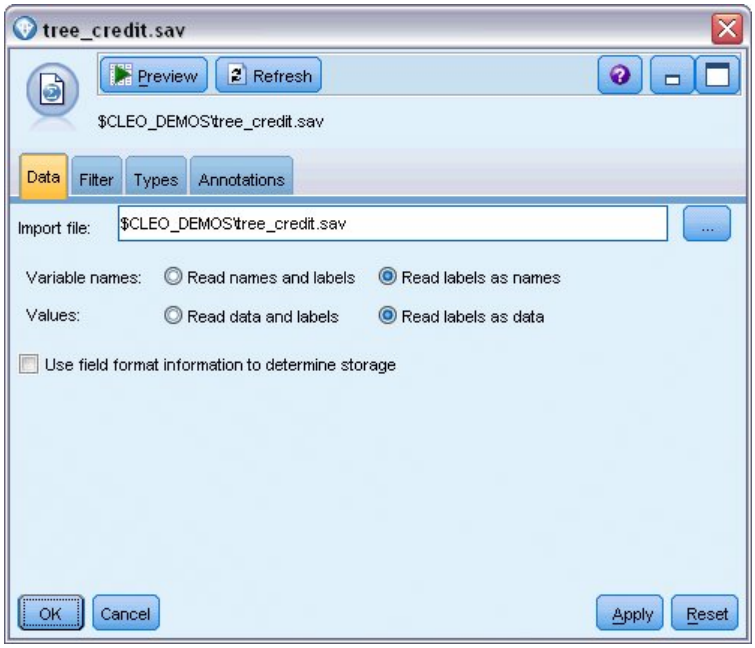


Figura 13. Lendo dados com um nó de origem Arquivo de Estatísticas

O nó Tipo especifica o **nível de medição** para cada campo. O nível de medição é uma categoria que indica o tipo de dados no campo. Nosso arquivo de dados de origem utiliza três níveis diferentes de medição.

Um campo **Contínuo** (como o campo *Idade*) contém valores numéricos contínuos, enquanto um campo **Nominal** (como o campo *Classificação de crédito*) tem dois ou mais valores distintos, por exemplo *Ruim*, *Bom* ou *Sem histórico de crédito*. Um campo **Ordinal** (como o campo *Nível de receita*) descreve dados com vários valores distintos que possuem uma ordem inerente - neste caso *Baixo*, *Médio* e *Alto*.

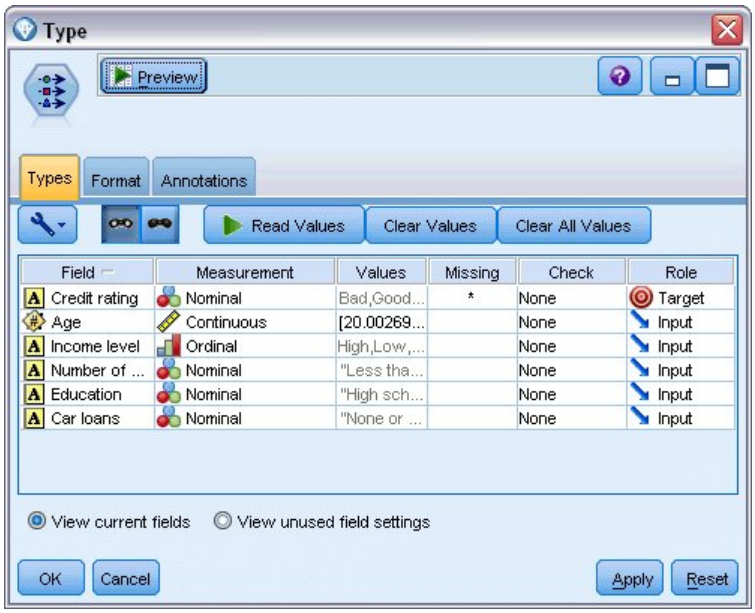


Figura 14. Configurando os campos de destino e de entrada com o nó Tipo

Para cada campo, o nó Tipo também especifica um **papel**, para indicar a função que cada campo atua na modelagem. função é configurada como *Destino* para o campo *Classificação de crédito*, que é o campo que indica se um determinado cliente está inadimplente ou não. Este é o **destino** ou o campo para o qual desejamos prever o valor.

A função é configurada como *Entrada* para os outros campos. Os campos de entrada às vezes são conhecidos como **preditores** ou campos cujos valores são usados pelo algoritmo de modelagem para prever o valor do campo de destino.

O nó de modelagem CHAID gera o modelo.

Na guia Campos no nó de modelagem, a opção **Usar papéis predefinidos** é selecionada, o que significa que o destino e as entradas serão utilizados conforme especificado no nó Tipo. Poderíamos alterar os papéis do campo neste momento, mas para este exemplo, eles serão usados no estado em que se encontram.

1. Clique na guia Opções de Criação.

Figura 15. Nó de modelagem CHAID, guia Campos

Aqui há várias opções que permite especificar o tipo de modelo que queremos construir.

Como queremos um modelo novo, vamos usar a opção **Construir novo modelo** padrão.

Também queremos um único modelo de árvore de decisão padrão sem quaisquer aprimoramentos, portanto, manteremos também a opção objetiva padrão **Construir uma árvore única**.

Embora seja possível, opcionalmente, ativar uma sessão de modelagem interativa que permite fazer um ajuste preciso do modelo, este exemplo simplesmente gera um modelo utilizando a configuração de modo padrão **Gerar modelo**.

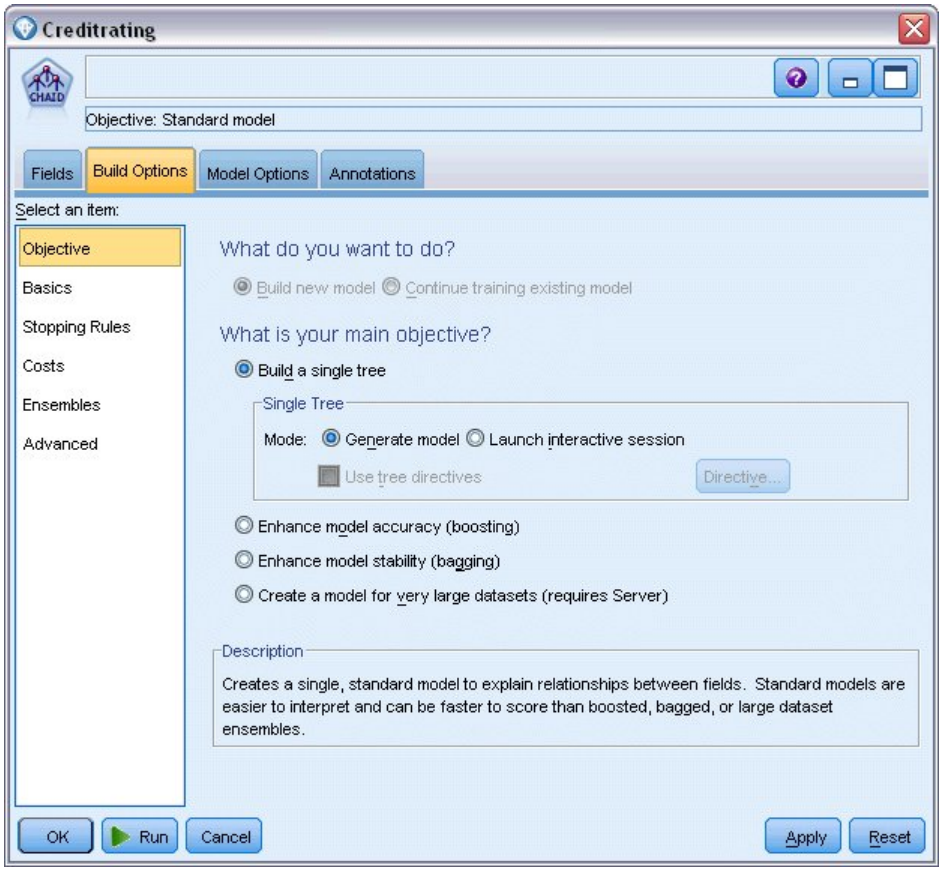


Figura 16. Nó de modelagem CHAID, guia Opções de Criação

Para esse exemplo, como queremos manter a árvore muito simples, vamos limitar o crescimento dela ao aumentar o número mínimo de casos para nós pais e filhos.

2. Na guia Opções de Criação, selecione **Regras de Parada** na área de janela do navegador à esquerda.
3. Selecione a opção **Usar valor absoluto**.
4. Configure o **Mínimo de registros na ramificação pai** para 400.
5. Configure o **Mínimo de registros na ramificação filha** para 200.

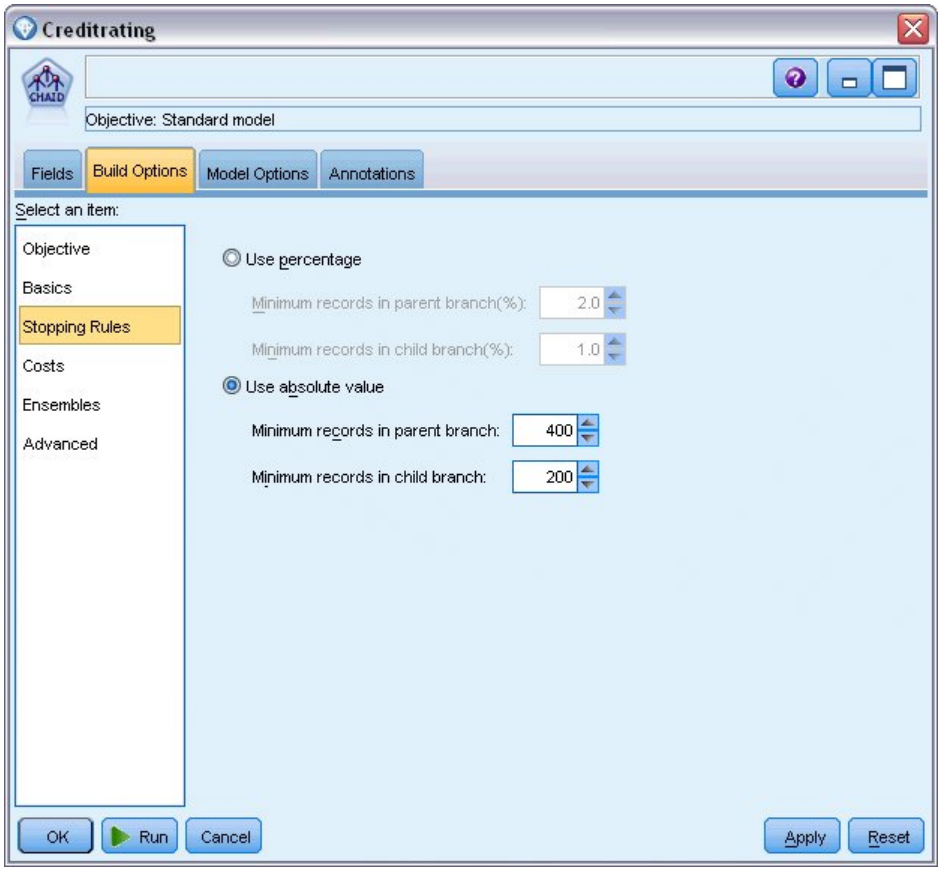


Figura 17. Configurando os critérios de parada para construção de árvore de decisão

Como é possível utilizar todas as outras opções padrão para este exemplo, clique em **Executar** para criar o modelo. (Como alternativa, clique com o botão direito no nó e escolha **Executar** no menu de contexto ou selecione o nó e escolha **Executar** no menu Ferramentas).

Procurando o Modelo

Quando a execução é concluída, o nugget do modelo é incluído na paleta Modelos no canto superior direito da janela do aplicativo, e também é colocado na tela de fluxo com uma ligação com o nó de modelagem a partir da qual ele foi criado. Para visualizar os detalhes do modelo, clique com o botão direito no nugget do modelo e escolha **Procurar** (na paleta de modelos) ou **Editar** (na tela).

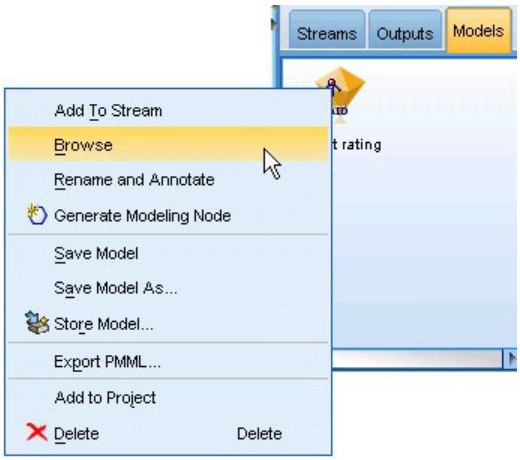


Figura 18. Paleta de Modelos

No caso do nugget CHAID, a guia Modelo exibe os detalhes na forma de um conjunto de regras -- essencialmente, uma série de regras que podem ser utilizadas para designar registros individuais para os nós filhos com base nos valores de diferentes campos de entrada.

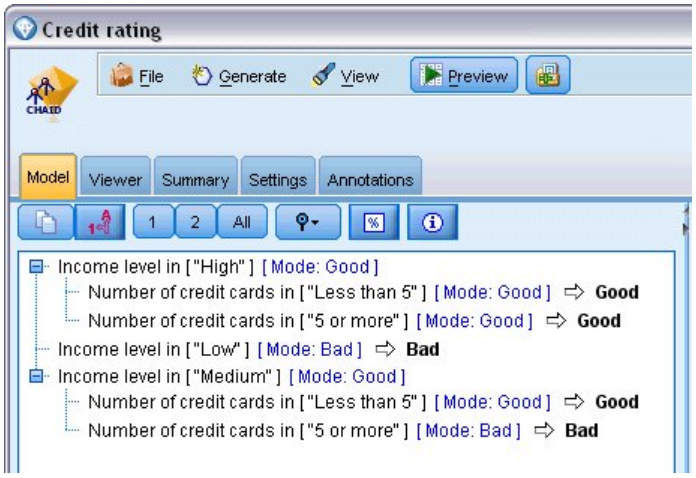


Figura 19. Nugget do modelo CHAID, conjunto de regras

Para cada nó terminal de árvore de decisão -- significando os nós de árvore que não são divididos ainda mais -- uma predição de *Bom* ou *Ruim* é retornada. Em cada caso, a predição é determinada pelo **Modo**, ou resposta mais comum, para registros que caírem nesse nó.

À direita do conjunto de regras, a guia Modelo exibe o gráfico Importância do Preditor, que mostra a importância relativa de cada preditor na estimativa do modelo. A partir disso, podemos ver que o *Nível de renda* é facilmente o mais significativo neste caso, e que o único outro fator significativo é *Número de cartões de crédito*.

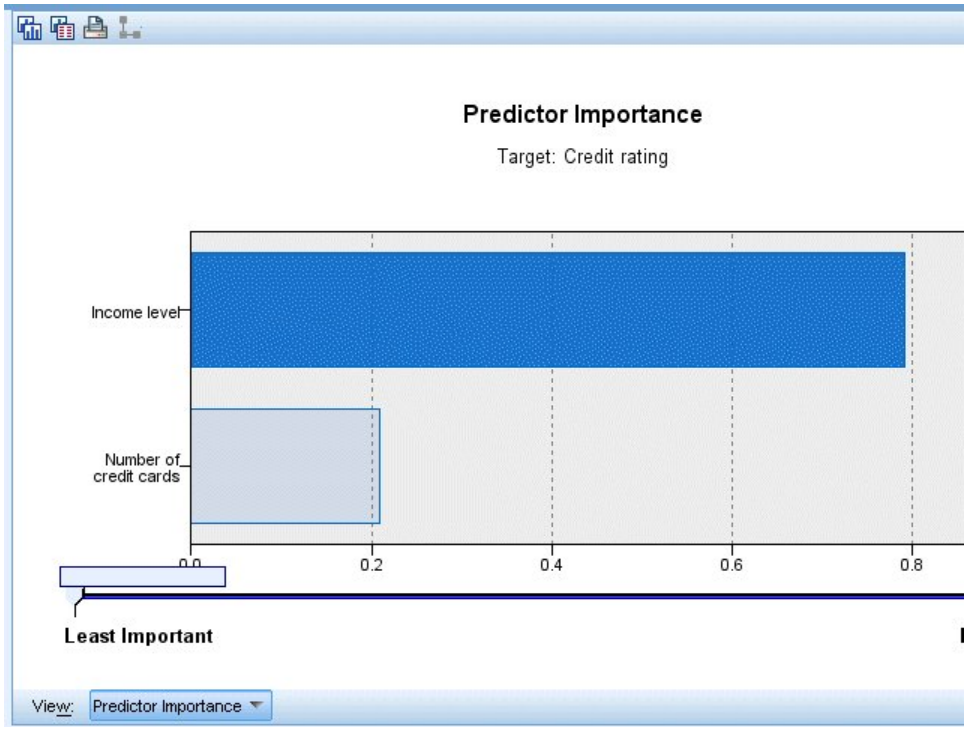


Figura 20. Gráfico de Importância do Preditor

A guia Visualizador no nugget do modelo exibe o mesmo modelo na forma de uma árvore, com um nó em cada ponto de decisão. Utilize os controles de Zoom na barra de ferramentas para aumentar o zoom em um nó específico ou diminuir o zoom para ver mais da árvore.

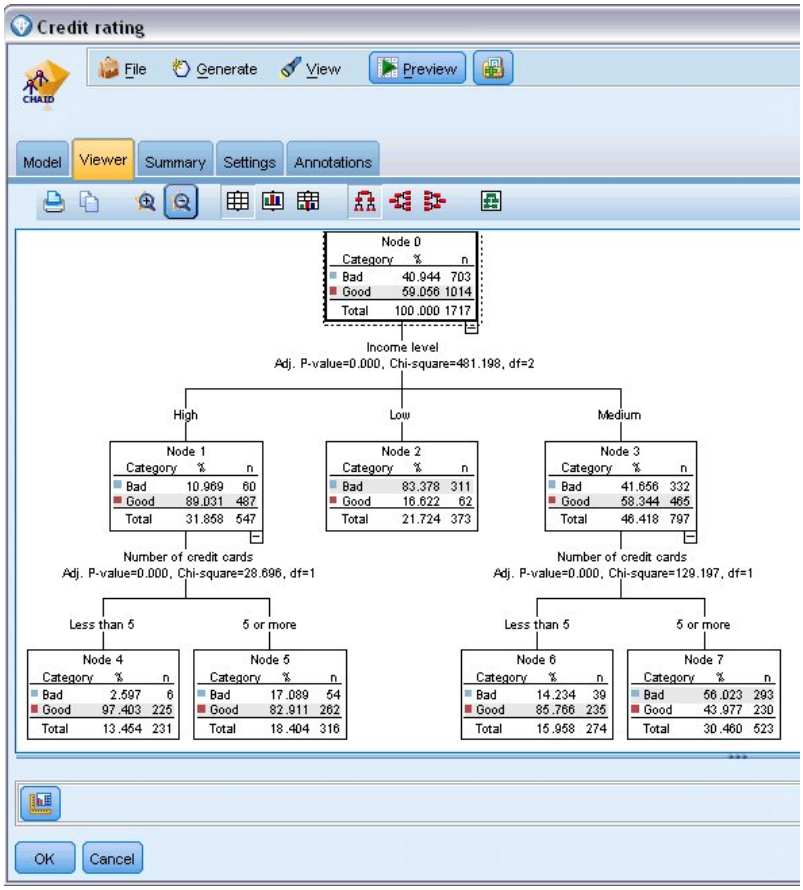


Figura 21. Guia Visualizador no nugget do modelo, com diminuir zoom selecionado

Observando a parte superior da árvore, o primeiro nó (Nó 0) fornece uma sumarização de todos os registros no conjunto de dados. Mais de 40% dos casos no conjunto de dados são classificados como um mau risco. Como esta é uma proporção muito alta, vamos verificar se a árvore pode dar dicas sobre quais fatores podem ser responsáveis.

Podemos ver que a primeira divisão é por *nível de renda*. Os registros em que o nível de renda estiver na categoria *Baixa* são designados ao Nó 2, e não é de surpreender que esta categoria contém a porcentagem mais alta de inadimplentes com empréstimos. É evidente que conceder empréstimos para clientes nesta categoria é altamente arriscado.

No entanto, como 16% dos clientes nesta categoria *não* estão realmente inadimplentes, a predição nem sempre estará correta. Nenhum modelo pode prever todas as respostas de maneira viável, mas um bom modelo deve nos permitir prever a resposta *mais provável* para cada registro com base nos dados disponíveis.

Da mesma forma, se olharmos os clientes de alta renda (Nó 1), vemos que a grande maioria (89%) representa um bom risco. No entanto, mais de 1 a cada 10 desses clientes também estiveram inadimplentes. Nós podemos refinar os critérios de empréstimo para minimizar o risco aqui?

Observe como o modelo dividiu esses clientes em duas subcategorias (Nós 4 e 5), com base no número de cartões de crédito que eles possuem. Para clientes de alta renda, se um empréstimo for concedido somente para clientes com menos de 5 cartões de crédito, poderemos aumentar a taxa de sucesso de 89% para 97% -- que é um resultado ainda mais satisfatório.

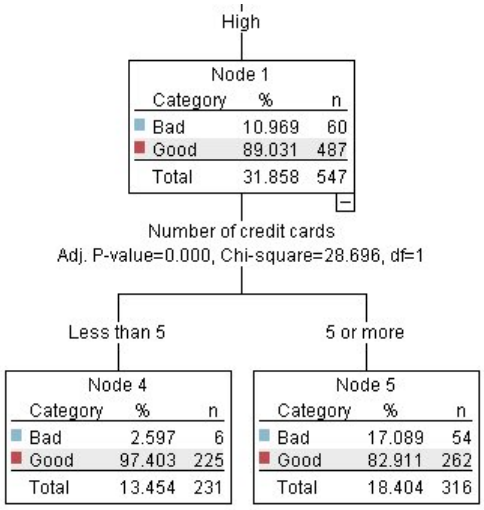


Figura 22. Visualização em árvore de clientes de alta renda

E quanto aos clientes na categoria Renda média (Nó 3)? Eles são muito mais igualmente divididos entre as classificações Bom e Ruim.

Mais uma vez, as subcategorias (Nós 6 e 7 nesse caso) podem nos ajudar. Desta vez, conceder empréstimo apenas para clientes com renda média com menos de 5 cartões de crédito aumenta a porcentagem de classificações Bom de 58% para 85%, que é uma melhoria significativa.

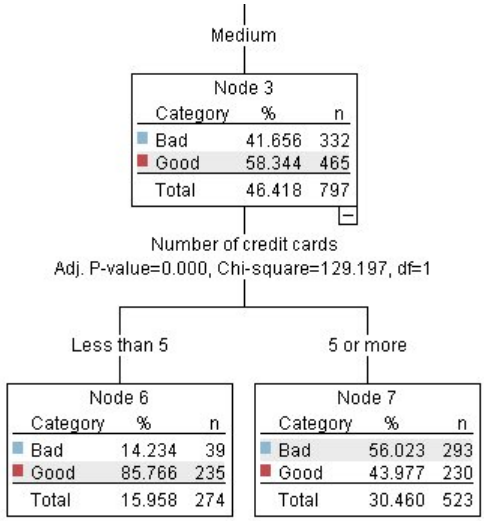


Figura 23. Visualização em árvore de clientes de renda média

Portanto, vimos que cada registro que é inserido nesse modelo é designado a um nó específico, e uma predição de *Bom* ou *Ruim* é designada com base na resposta mais comum para esse nó.

Esse processo de atribuição de previsões a registros individuais é conhecido como **escoragem**. Ao escorar os mesmos registros usados para estimar o modelo, podemos avaliar precisamente como será o desempenho desse modelo nos dados de treinamento – os dados para os quais sabemos o resultado. Vamos ver como fazer isso.

Avaliando o Modelo

Nós temos procurado o modelo para entender como a escoragem funciona. Mas para avaliar *a precisão* com que funciona, precisamos pontuar alguns registros e comparar as respostas previstas pelo modelo com os resultados reais. Iremos escorar os mesmos registros que foram utilizados para estimar o modelo, permitindo comparar as respostas observadas e preditas.

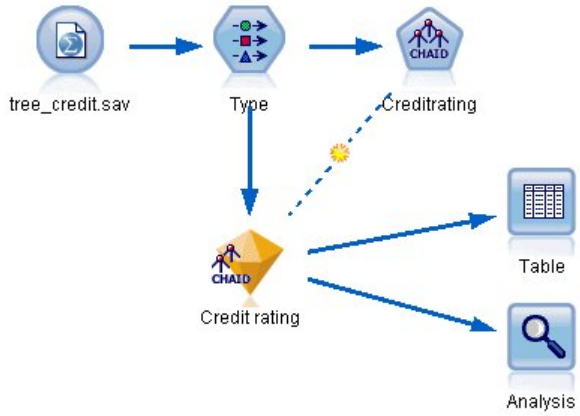
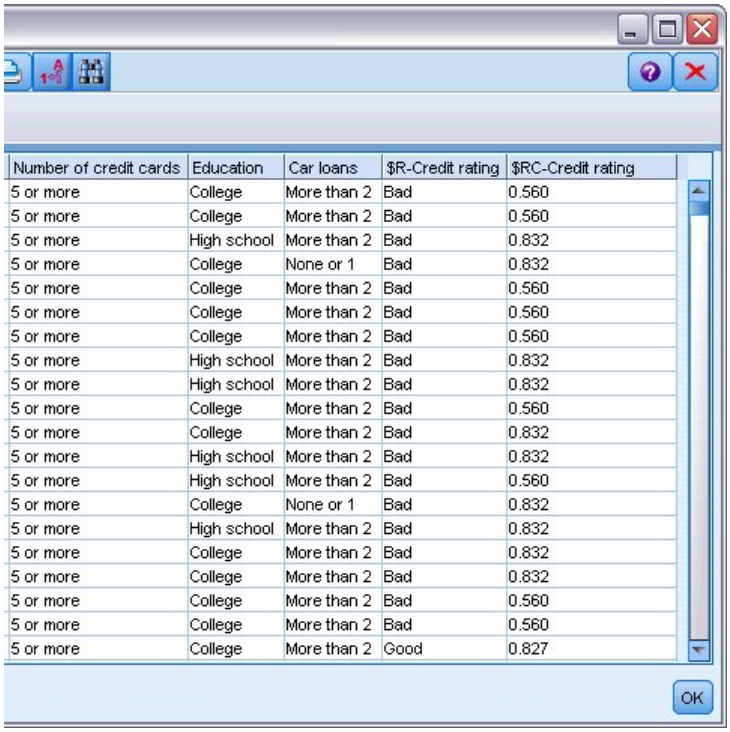


Figura 24. Anexado o nugget do modelo aos nós de saída para avaliação de modelo

1. Para ver as escoragens ou predições, anexe o nó Tabela ao nugget do modelo, clique duas vezes no nó Tabela e clique em **Executar**.

   A tabela exibe as pontuações previstas em um campo denominado *$R-Credit rating*, que foi criado pelo modelo. Podemos comparar esses valores com o campo de *classificação de crédito* original que contém as respostas reais.

   Por convenção, os nomes dos campos gerados durante a escoragem baseiam-se no campo de destino, mas com um prefixo padrão. Os prefixos *$G* e *$GE* são gerados pelo Modelo de Linear Generalizado, *$R* é o prefixo usado para a predição gerada pelo modelo CHAID neste caso, *$RC* é para valores de confiança, *$X* geralmente é gerado usando um conjunto, e *$XR*, *$XS* e *$XF* são usados como prefixos nos casos em que o campo de destino é um campo Contínuo, Categórico, Conjunto ou Sinalizador, respectivamente. Tipos de modelo diferentes utilizam conjuntos diferentes de prefixos. Um **valor de confiança** é a estimação do próprio modelo do grau de precisão de cada valor predito, em uma escala de 0,0 a 1,0.



Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	College	None or 1	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	High school	More than 2	Bad	0.560
5 or more	College	None or 1	Bad	0.832
5 or more	High school	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.832
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Bad	0.560
5 or more	College	More than 2	Good	0.827

Figura 25. Tabela mostrando escoragens geradas e valores de confiança

Conforme esperado, o valor predito corresponde às respostas reais para muitos registros, mas não para todos. O motivo para isso é que cada nó terminal CHAID possui uma combinação de respostas. A

predição corresponde à *mais comum*, mas estará errada para todas as outras naquele nó. (Lembre-se da minoria de 16% de clientes de baixa renda que não estiveram inadimplentes).

Para evitar isso, poderíamos continuar dividindo a árvore em ramos cada vez menores, até que cada nó fosse 100% puro - todos *bons* ou *ruins*, sem respostas mistas. No entanto, esse modelo seria extremamente complicado e provavelmente não generalizaria tão bem para os demais conjuntos de dados.

Para saber exatamente quantas previsões estão corretas, poderíamos ler a tabela e calcular o número de registros em que o valor do campo previsto *$R-Credit rating* corresponde ao valor de *Credit rating*. Felizmente, há outra maneira muito mais fácil que é utilizar o nó Análise que faz todo esse processo automaticamente.

2. Conecte o nugget do modelo ao nó Análise.
3. Clique duas vezes no nó Análise e clique em **Executar**.

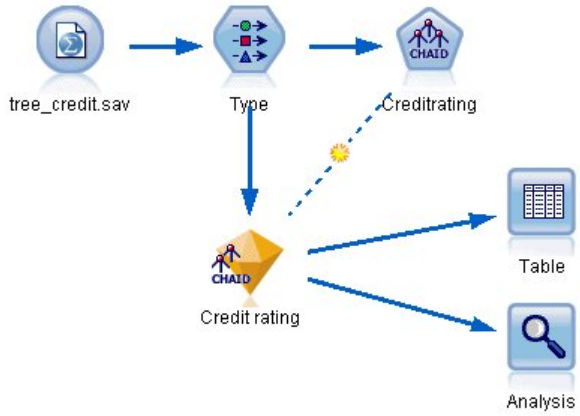


Figura 26. Anexando um nó de análise

A análise mostra que, para 1899 de 2464 registros -- mais de 77% -- o valor predito pelo modelo correspondeu à resposta real

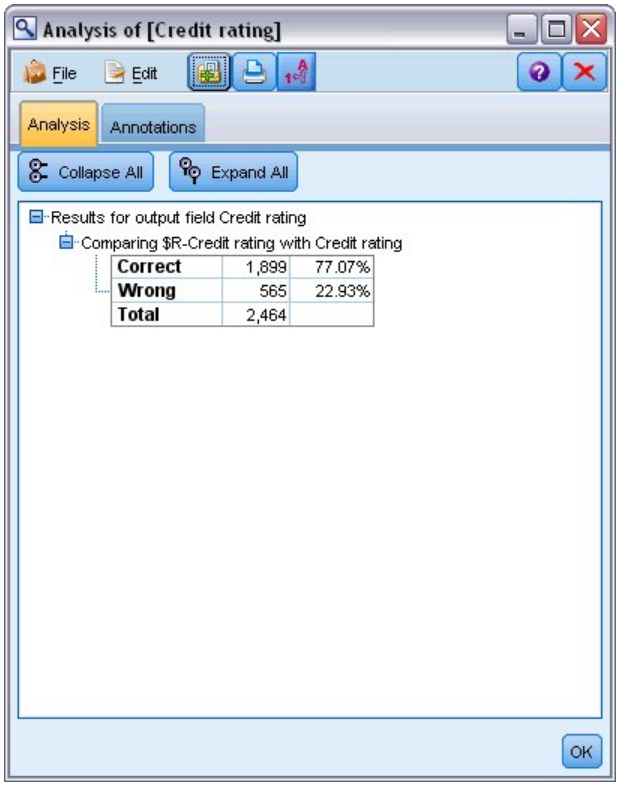


Figura 27. Resultados da análise comparando respostas observadas e preditas

Esse resultado é limitado pelo fato de que os registros que estão sendo escorados são os mesmos utilizados para estimar o modelo. Em uma situação real, é possível utilizar um nó Partição para dividir os dados em amostras separadas para treinamento e avaliação.

Ao utilizar uma partição de amostra para gerar o modelo e outra amostra para testá-lo, é possível obter uma indicação muito melhor do quão bem ele será generalizado para outros conjuntos de dados.

O nó Análise permite testar o modelo com relação aos registros para os quais nós já sabemos o resultado real. O próximo passo ilustra como é possível utilizar o modelo para escorar registros para os quais não sabemos o resultado. Por exemplo, isso pode incluir pessoas que não forem atualmente clientes de um banco, mas que são possíveis alvos de receberem um email promocional.

Escoragem de registros

Anteriormente, nós escoramos os mesmos registros utilizados para estimar o modelo para avaliar o nível de precisão do modelo. Agora vamos ver como escorar um conjunto diferente de registros a partir daqueles utilizados para criar o modelo. Este é o objetivo de modelagem com um campo de destino: Registros de estudo para os quais você sabe o resultado, para identificar padrões que permitirão prever resultados que você ainda não sabe.

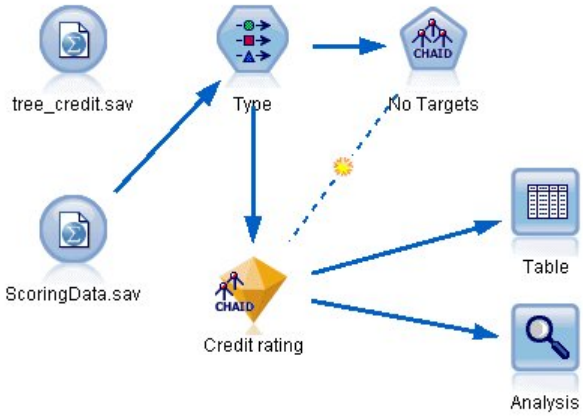


Figura 28. Anexando novos dados para escoragem

É possível atualizar o nó de origem Arquivo de Estatísticas para apontar para um arquivo de dados diferente, ou incluir um novo nó de origem que lê nos dados que você deseja escorar. De qualquer maneira, o novo conjunto de dados deverá conter os mesmos campos de entrada utilizados pelo modelo (*Idade*, *Nível de renda*, *Educação*, e assim por diante), mas não o campo de destino *Classificação de Crédito*.

Como alternativa, é possível incluir o nugget do modelo em qualquer fluxo que inclua os campos de entrada esperados. Independentemente de ler a partir de um arquivo ou de um banco de dados, o tipo de origem não importa desde que os nomes e tipos de campos correspondam aos utilizados pelo modelo.

Você também poderia salvar o nugget de modelo como um arquivo separado, ou exportar o modelo em formato PMML para uso com outros aplicativos que suportam este formato, ou armazenar o modelo em um repositório IBM SPSS Serviços de colaboração e implantação , que oferece implantação, pontuação e gerenciamento de modelos de grande escala.

Independentemente da infraestrutura utilizada, o próprio modelo funciona da mesma maneira.

Resumo

Este exemplo demonstra os passos básicos para criar, avaliar e escorar um modelo.

- O nó de modelagem estima o modelo ao estudar os registros para os quais o resultado é conhecido, e cria um nugget do modelo. Às vezes isso é referido como treinamento do modelo.
- O nugget do modelo pode ser incluído em qualquer fluxo com os campos esperados para escorar registros. Ao escorar os registros para os quais você já sabe o resultado (como clientes existentes), é possível avaliar o seu grau de desempenho.
- Quando estiver satisfeito com o desempenho do modelo, será possível escorar novos dados (como clientes esperados) para prever como eles responderão.
- Os dados usados para treinar ou estimar o modelo podem ser referidos como dados de analítica ou históricos, e os dados de escoragem também podem ser referidos como os dados operacionais.

Capítulo 4. Modelagem automatizada para um alvo de sinalização

Modelagem de resposta do cliente (classificador automático)

O nó Auto Classifier permite que você crie e compare automaticamente vários modelos diferentes para qualquer sinalização (como se um determinado cliente deve ou não ficar inadimplente em um empréstimo ou responder a uma determinada oferta) ou destinos nominais (set). Neste exemplo vamos procurar um resultado de sinalização (sim ou não). Dentro de um fluxo relativamente simples, o nó gera e classifica um conjunto de modelos de candidatos, escolhe aqueles que executam o melhor, e os combina em um modelo único agregado (Ensembled). Essa abordagem combina a facilidade de automação com os benefícios de combinar múltiplos modelos, que muitas vezes geram previsões mais precisas do que podem ser obtidas com qualquer modelo.

Este exemplo é baseado em uma empresa fictícia que deseja obter resultados mais lucrativos combinando a oferta certa para cada cliente.

Essa abordagem enfatiza os benefícios da automação. Para um exemplo semelhante que usa um alvo contínuo (faixa numérica), veja Valores da propriedade (Auto Numeric).

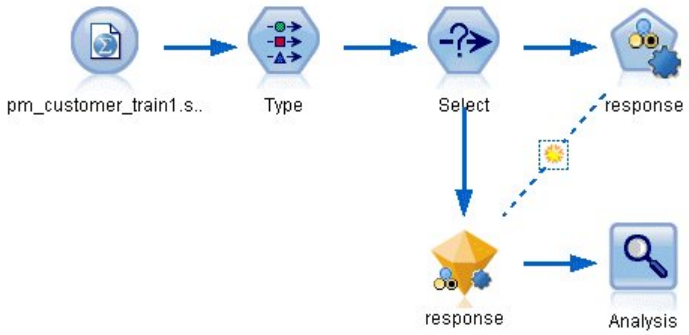


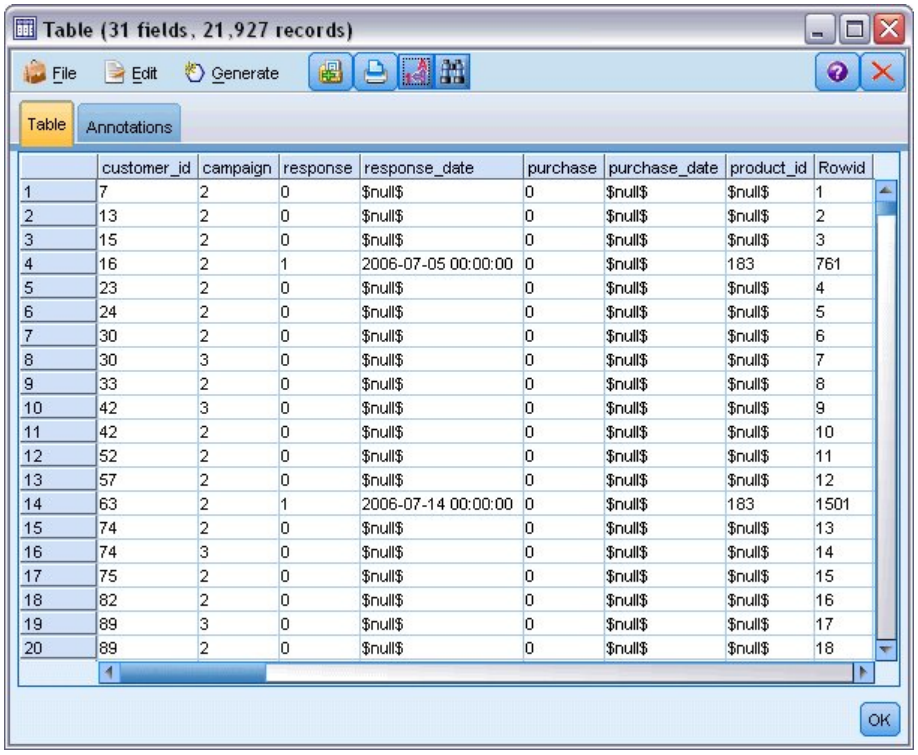
Figura 29. Fluxo de amostra de Auto Classifier

Este exemplo usa o fluxo *pm_binaryclassifier.str*, instalado na pasta Demo em *streams*. O arquivo de dados usado é *pm_customer_train1.sav*.. Consulte o tópico “Dados históricos” na página 33 para obter mais informações.

Dados históricos

O arquivo *pm_customer_train1.sav* tem dados históricos rastreando as ofertas feitas a clientes específicos em campanhas passadas, conforme indicado pelo valor do campo *campanha* . O maior número de registros cai na campanha de *conta premium*.

Os valores do campo da *campanha* são codificados como números inteiros nos dados (por exemplo *2 = conta Premium*). Posteriormente, você definirá rótulos para esses valores que podem ser usados para fornecer uma saída mais significativa.



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

Figura 30. Dados sobre promoções anteriores

O arquivo também inclui um campo *resposta* que indica se a oferta foi aceita (0 = *no*, e 1 = *yes*). Este será o **campo de destino**, ou valor, que você deseja prever. Vários campos contendo informações demográficas e financeiras sobre cada cliente também estão incluídos. Eles podem ser usados para construir ou "treinar" um modelo que prevê taxas de resposta para indivíduos ou grupos com base em características como renda, idade ou número de transações por mês.

Construindo o Fluxo

1. Adicione um nó de origem do Arquivo Estatísticas apontando para *pm_customer_train1.sav*, localizado na pasta *Demos* de sua instalação IBM SPSS Modelador . (Você pode especificar `$CLEO_DEMOS/` no caminho de arquivo como um atalho para referencia esta pasta. Observe que uma barra-em vez de uma barra invernada-deve ser usada no caminho, como mostrado.)

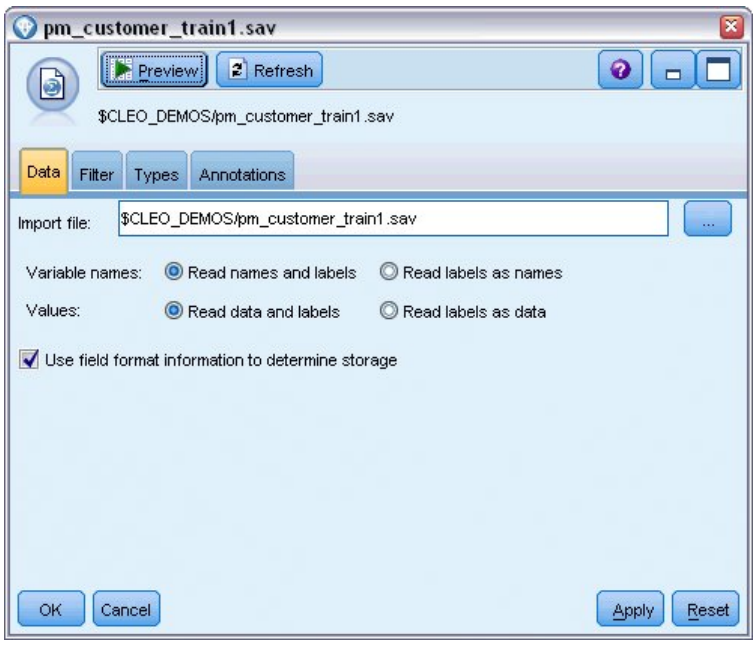


Figura 31. Leitura nos dados

2. Inclua um nó Tipo e selecione *resposta* como o campo de destino (Role = **Destino**). Configure a Medição para este campo para **Bandeira**.

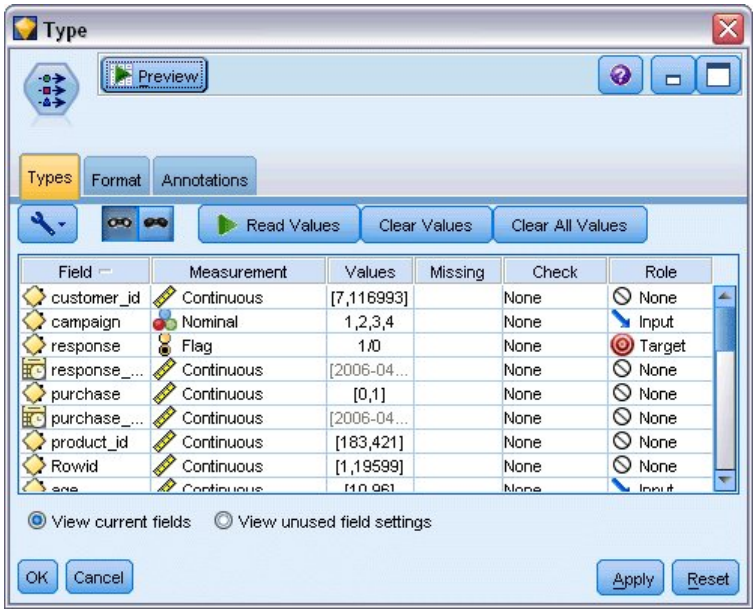


Figura 32. Configurando o nível de medição e a função

3. Configure a função para **Nenhum** para os campos a seguir: *customer_id*, *campanha*, *response_date*, *compra*, *purchase_date*, *product_id*, *Rowid* e *X_random*. Esses campos serão ignorados quando você estiver construindo o modelo.
4. Clique no botão **Valores de leitura** no nó Tipo para ter certeza de que os valores são instanciados.

   Como vimos anteriormente, nossos dados de origem incluem informações sobre quatro campanhas diferentes, cada uma direcionada a um tipo diferente de conta de cliente. Essas campanhas são codificadas como inteiros nos dados, de modo que seja mais fácil lembrar qual tipo de conta cada número inteiro representa, vamos definir rótulos para cada uma.

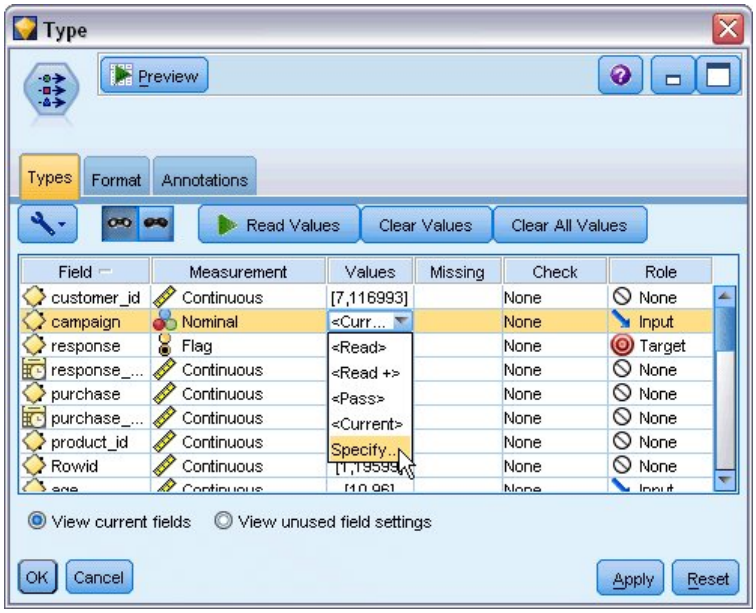


Figura 33. Optando por especificar valores para um campo

5. Na linha para o campo **campanha** , clique na entrada na coluna **Vales** .
6. Escolha **Especificar** na lista suspensa.

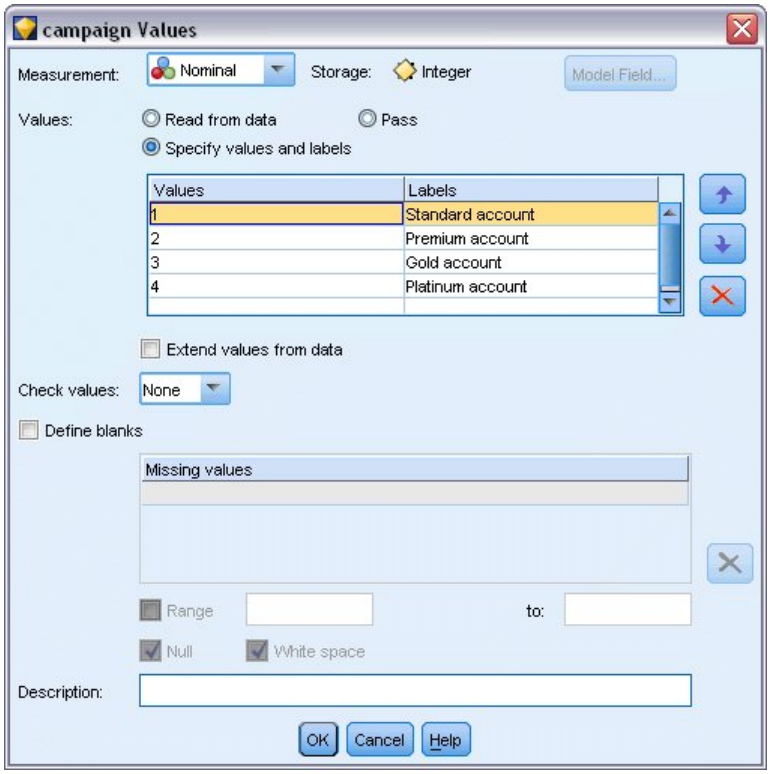
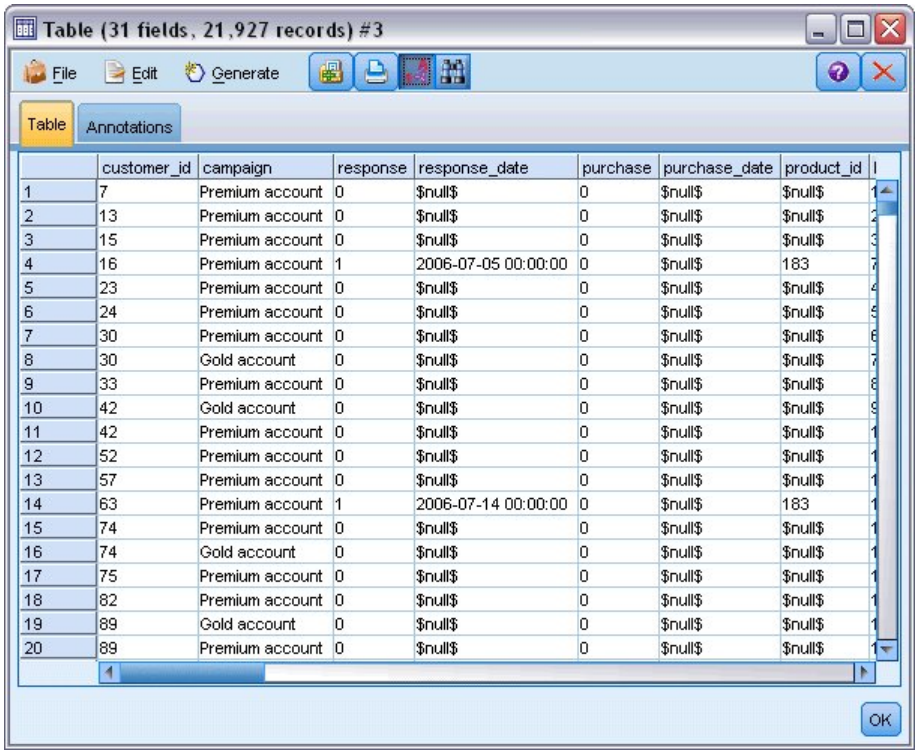


Figura 34. Definindo rótulos para os valores de campo

7. Na coluna **Labels** , digite as etiquetas conforme mostrado para cada um dos quatro valores do campo **campanha** .
8. Clique em **OK**.

Agora é possível exibir os rótulos em janelas de saída em vez dos números inteiros.



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id
1	7	Premium account	0	$null$	0	$null$	$null$
2	13	Premium account	0	$null$	0	$null$	$null$
3	15	Premium account	0	$null$	0	$null$	$null$
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183
5	23	Premium account	0	$null$	0	$null$	$null$
6	24	Premium account	0	$null$	0	$null$	$null$
7	30	Premium account	0	$null$	0	$null$	$null$
8	30	Gold account	0	$null$	0	$null$	$null$
9	33	Premium account	0	$null$	0	$null$	$null$
10	42	Gold account	0	$null$	0	$null$	$null$
11	42	Premium account	0	$null$	0	$null$	$null$
12	52	Premium account	0	$null$	0	$null$	$null$
13	57	Premium account	0	$null$	0	$null$	$null$
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183
15	74	Premium account	0	$null$	0	$null$	$null$
16	74	Gold account	0	$null$	0	$null$	$null$
17	75	Premium account	0	$null$	0	$null$	$null$
18	82	Premium account	0	$null$	0	$null$	$null$
19	89	Gold account	0	$null$	0	$null$	$null$
20	89	Premium account	0	$null$	0	$null$	$null$

Figura 35. Exibindo as etiquetas de valor de campo

9. Anexe um nó de tabela ao nó Tipo.
10. Abra o nó da Tabela e clique em **Executar**.
11. Na janela de saída, clique no botão **Exibir etiquetas e etiquetas** botão da barra de ferramentas para exibir os rótulos.
12. Clique em **OK** para fechar a janela de saída.

Embora os dados incluam informações sobre quatro campanhas diferentes, você concentrará a análise em uma campanha por vez. Uma vez que o maior número de registros cai sob a campanha da conta Premium (codificada *campaign=2* nos dados), você pode usar um nó Select para incluir apenas estes registros no fluxo.

Figura 36. Selecionando registros para uma única campanha

Gerando e comparando modelos

1. Conecte um nó do Auto Classifier, e selecione **Precisão geral** como a métrica usada para classificar modelos.
2. Defina o **Número de modelos a serem usados** como 3. Isso significa que os três melhores modelos serão construídos quando você executar o nó.

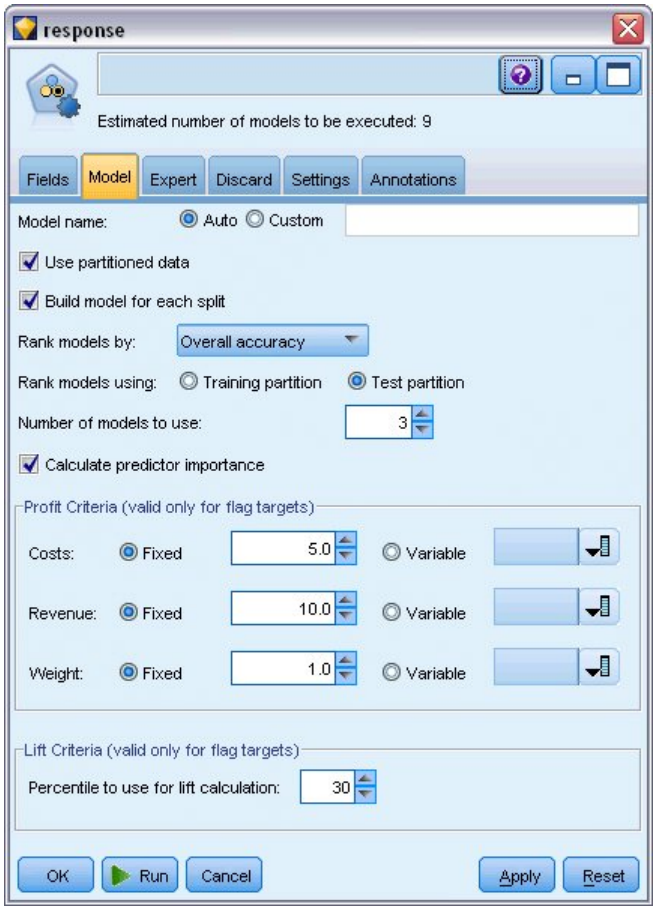


Figura 37. Guia Modelo de nó do Classificador Auto

Na guia Expert você pode escolher de até 11 algoritmos de modelos diferentes.

3. Desmarque os tipos de modelo **Discriminante** e **SVM**. (Esses modelos demoram mais para treinar nesses dados, então deselegá-los vai acelerar o exemplo. Se você não se importar em esperar, fique à vontade para deixá-los selecionados.)

   Como você configurou **Número de modelos a serem usados** a 3 na guia Modelo, o nó calculará a precisão dos nove algoritmos restantes e construirá um nugget de modelo único contendo os três mais precisos.

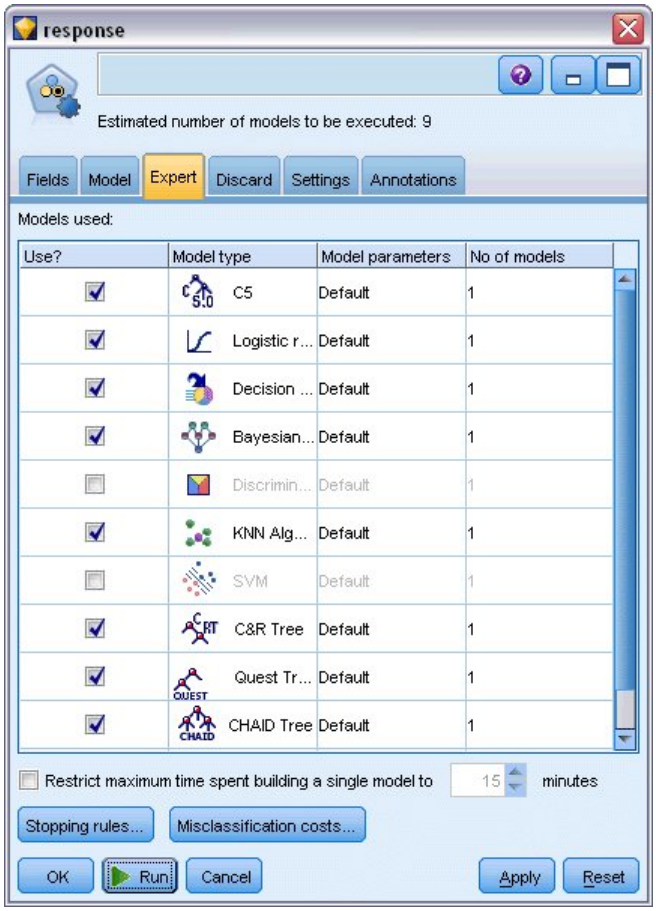


Figura 38. Guia Expert do nó Classificador automático

4. Na guia Configurações, para o método ensemble, selecione **Voo ponderado por Confiança**. Isso determina como uma única pontuação agregada é produzida para cada registro.

   Com a votação simples, se dois de três modelos predizerem *yes*, então o *yes* ganha por uma votação de 2 a 1. No caso de votação com ponderação de confiança, os votos são ponderados com base no valor de confiança de cada predição. Assim, se um modelo prediz *no* com uma confiança mais alta do que as duas predições *yes* combinadas, então *no* vence.

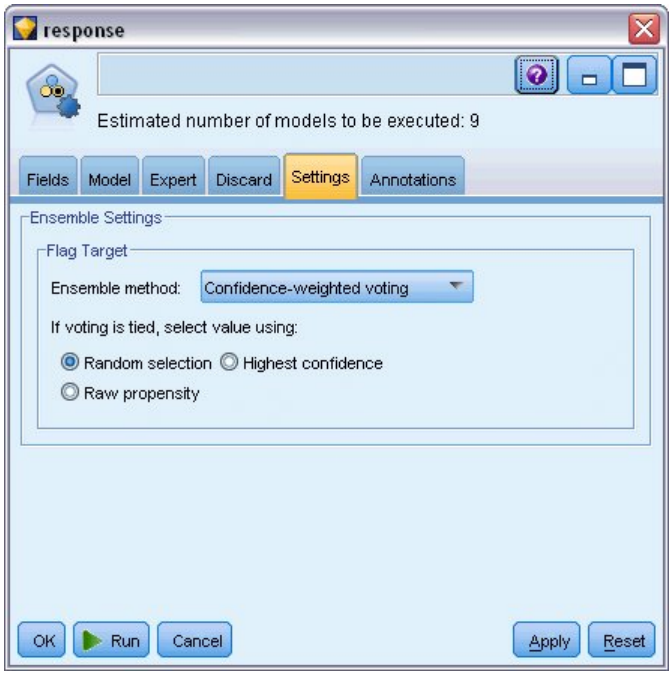


Figura 39. Nó do Classificador Automático: guia Configurações

5. Clique em **Executar** .

Após alguns minutos, o nugget de modelo gerado é construído e colocado na tela, e na paleta de Models no canto superior direito da janela. Você pode navegar no nugget do modelo, ou salvar ou implantá-lo em várias outras formas.

Abra o nugget modelo; ele lista detalhes sobre cada um dos modelos criados durante a execução. (Em uma situação real, em que centenas de modelos podem ser criados em um grande conjunto de dados, isso pode levar muitas horas.) Consulte Figura 29 na página 33.

Se você deseja explorar mais adiante qualquer um dos modelos individuais, você pode clicar duas vezes em um ícone de nugget de modelo na coluna **Modelo** para realizar a pesquisa e navegar nos resultados do modelo individual; a partir daí você pode gerar nós de modelagem, nuggets de modelo ou gráficos de avaliação. Na coluna **Gráfico** , é possível clicar duas vezes em uma miniatura para gerar um gráfico de tamanho médio.

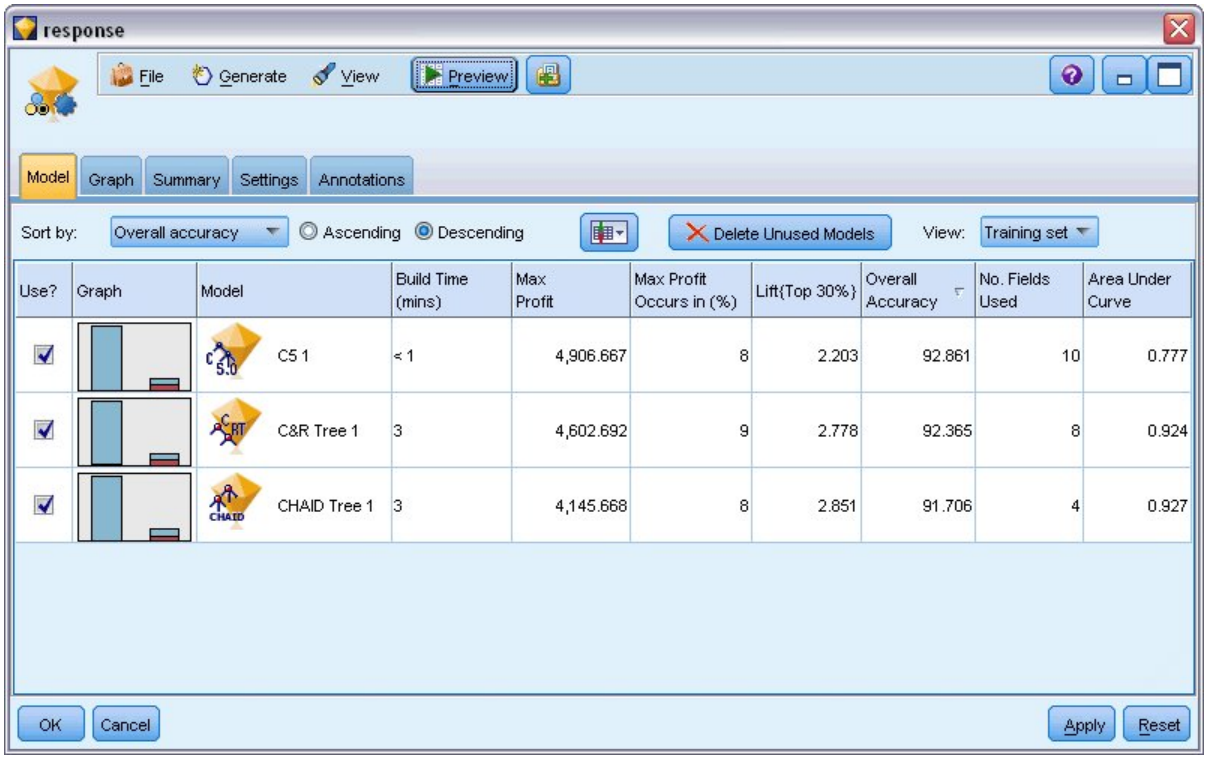


Figura 40. Resultados do classificador automático

Por padrão, os modelos são classificados com base na exatidão geral, pois esta foi a medida que você selecionou na aba Modelo de nó do Classificador Auto. O modelo C51 se classifica melhor por esta medida, mas os modelos C & R Tree e CHAID são quase tão precisos.

Você pode classificar em uma coluna diferente clicando no cabeçalho para essa coluna, ou você pode escolher a medida desejada a partir da lista suspensa **Classificar por** na barra de ferramentas.

Com base nesses resultados, você decide usar todos os três desses modelos mais precisos. Ao combinar predições a partir de vários modelos, as limitações em modelos individuais podem ser evitadas, resultando em uma precisão geral mais alta.

No **Use?** coluna, selecione os modelos C51, C & R Tree e CHAID.

Anexar um nó de Análise (paleta de saída) após o nugget modelo. Clique com o botão direito do mouse sobre o nó da Análise e escolha **Executar** para executar o fluxo.

A pontuação agregada gerada pelo modelo combinado é mostrada em um campo denominado *$XF-response*. Quando medido com relação aos dados de treinamento, o valor predito corresponde à resposta real (conforme registrado no campo *resposta* original) com uma precisão geral de 92.82%.

Embora não seja tão preciso quanto o melhor dos três modelos individuais neste caso (92.86% para C51), a diferença é muito pequena para ser significativa Em termos gerais, um modelo combinado normalmente terá mais probabilidade de ter um bom desempenho quando aplicado a conjuntos de dados diferentes dos dados de treinamento.

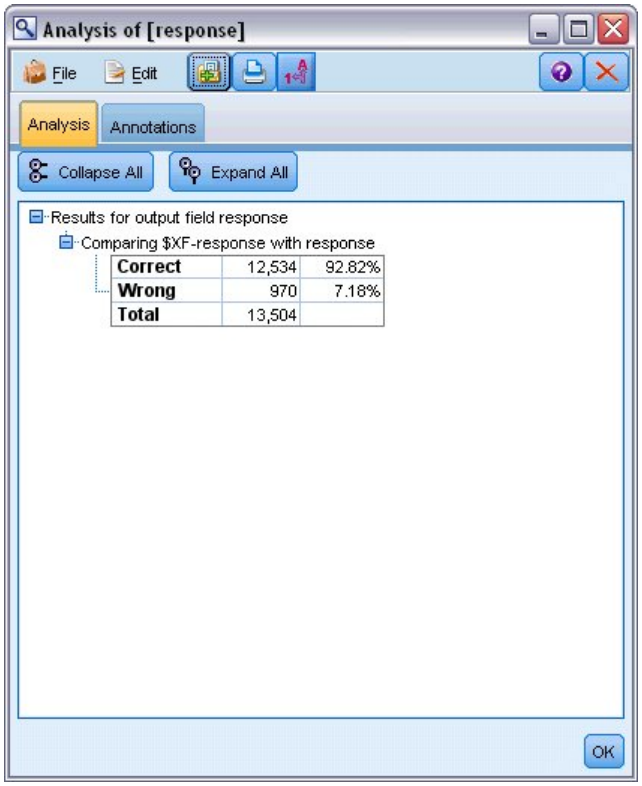


Figura 41. Análise dos três modelos combinados

Resumo

Para resumir, você usou o nó Auto Classifier para comparar uma série de modelos diferentes, usou os três modelos mais precisos e os adicionou ao fluxo dentro de uma nugget modelo de Auto Classifier ensembled.

- Com base na precisão geral, os modelos C51, C & R Tree e CHAID apresentaram melhor desempenho nos dados de treinamento.
- O modelo combinado teve um desempenho quase tão bom quanto o melhor dos modelos individuais e pode ter um desempenho melhor quando aplicado a outros conjuntos de dados. Se seu objetivo é automatizar o processo tanto quanto possível, esta abordagem permite que você obtenha um modelo robusto na maioria das circunstâncias, sem ter que se aprofundar nas especificações de qualquer modelo.

Capítulo 5. Modelagem automatizada para um destino contínuo

Valores de propriedade (Previsor contínuo automático)

O Auto Numérico permite criar e comparar automaticamente diferentes modelos para resultados contínuos (faixa numérica), como prever o valor tributável de um imóvel. Com um único nó, é possível estimar e comparar um conjunto de modelos candidatos e gerar um subconjunto de modelos para análise posterior. O nó funciona da mesma maneira que o nó classificador automático, mas para alvos contínuos em vez de sinalizadores ou nominais.

O nó combina o melhor dos modelos candidatos em um único nugget do modelo agregado (Ensembled). Essa abordagem combina a facilidade de automação com os benefícios de combinar múltiplos modelos, que muitas vezes geram previsões mais precisas do que podem ser obtidas com qualquer modelo.

Este exemplo se concentra em um município fictício responsável por ajustar e avaliar os impostos imobiliários. Para fazer isso com mais precisão, eles construirão um modelo que prevê os valores das propriedades com base no tipo de construção, vizinhança, tamanho e outros fatores conhecidos.

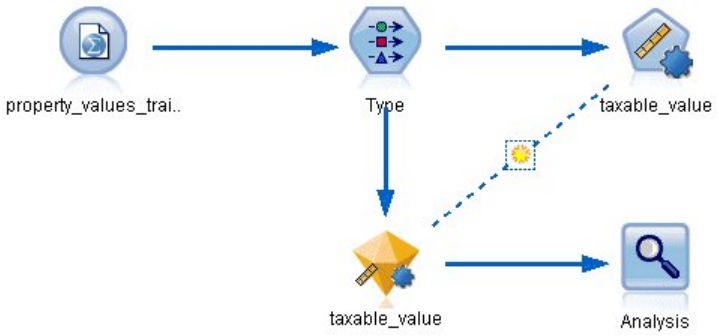


Figura 42. Fluxo De Amostra Numérico Automático

Este exemplo usa o fluxo `property_values_numericpredictor.str`, instalado na pasta Demos sob `streams`. O arquivo de dados utilizado é `property_values_train.sav`. Veja o tópico "Pasta Demos" na página 4 para obter mais informações.

Dados de treinamento

O arquivo de dados inclui um campo denominado *taxable_value*, que é o **campo de destino** ou valor que você deseja prever. Os outros campos contêm informações como vizinhança, tipo de construção e volume interior e podem ser usados como preditores.

Nome do campo	Rótulo
proprity_id	ID da propriedade
ambiente	Área dentro da cidade
building_type	Tipo de edifício
year_construído	Ano de construção
volume_interior	Volume do interior
volume_outro	Volume de garagem e edifícios extras
tamanho do lote	Tamanho do lote

Nome do campo	Rótulo
taxable_value (valor tributável)	Valor tributável

Um arquivo de dados de pontuação denominado *property_values_score.sav* também é incluído na pasta Demos Ele contém os mesmos campos mas sem o campo *taxable_value* . Após os modelos de treinamento usando um dataset onde o valor tributável é conhecido, é possível pontuar registros onde esse valor ainda não é conhecido.

Construindo o Fluxo

1. Inclua um nó de origem de arquivo Estatísticas apontando para *property_values_train.sav*, localizado na pasta *Demos* de sua instalação do IBM SPSS Modelador (Você pode especificar `$CLEO_DEMOS/` no caminho de arquivo como um atalho para referencia esta pasta. Observe que uma barra-em vez de uma barra invernada-deve ser usada no caminho, como mostrado.)

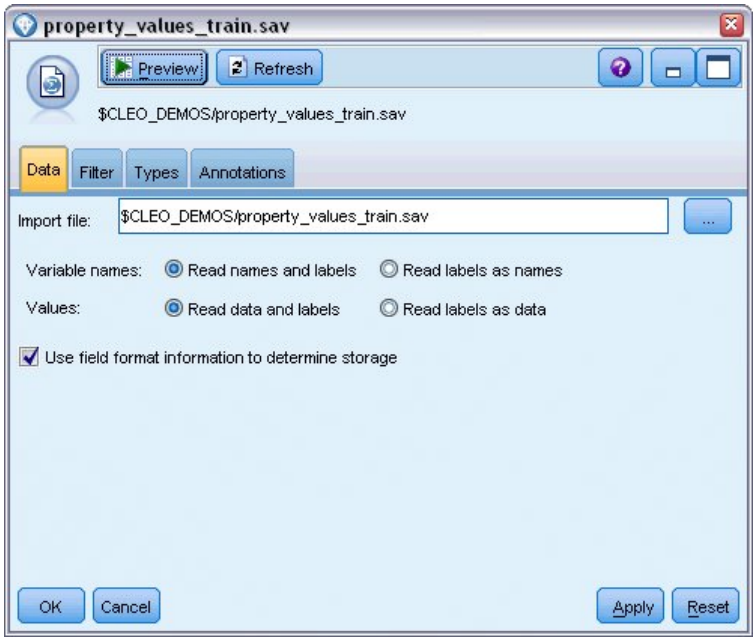


Figura 43. Leitura nos dados

2. Inclua um nó Tipo e selecione *taxable_value* como o campo de destino (Role = **Target**). Função deve ser configurada como **Entrada** para todos os outros campos, indicando que eles serão usados como preditores.

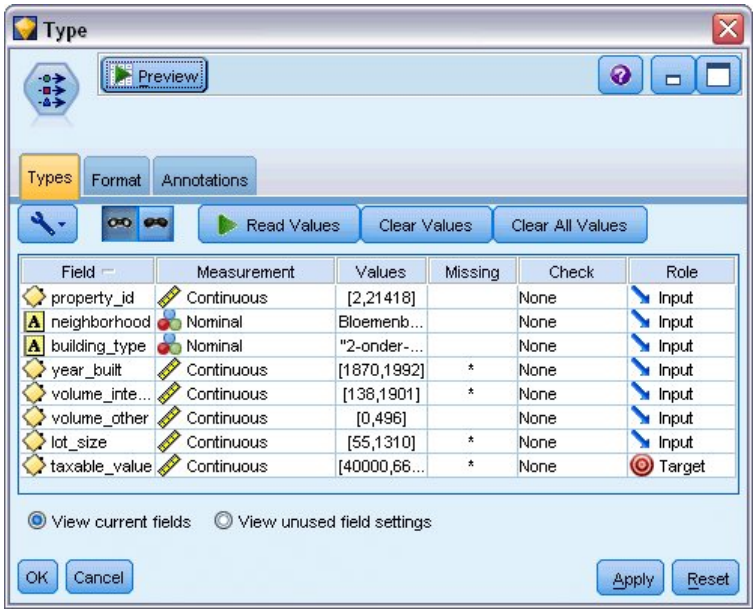


Figura 44. Configurando o campo de destino

3. Conecte um nó Numeric automático, e selecione **Correlação** como a métrica usada para classificar modelos.
4. Defina o **Número de modelos a serem usados** como 3. Isso significa que os três melhores modelos serão construídos quando você executar o nó.

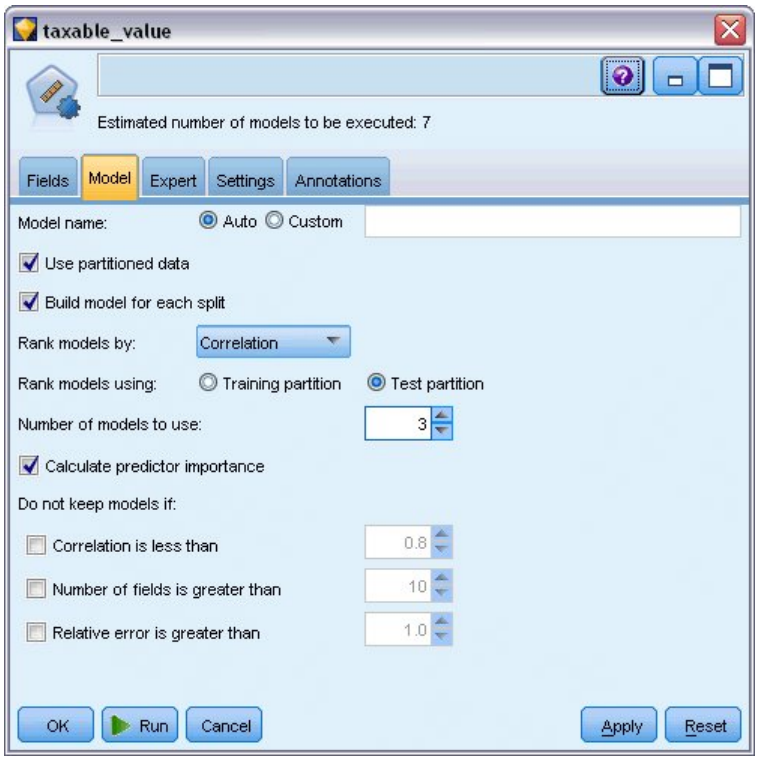


Figura 45. Guia Modelo de nó Numérico automático

5. Na guia Expert, deixe as configurações padrão no lugar; o nó irá estimar um modelo único para cada algoritmo, para um total de sete modelos. (Como alternativa, é possível modificar essas configurações para comparar várias variantes para cada tipo de modelo.)

   Como você configura **Número de modelos a serem usados** a 3 na guia Modelo, o nó calculará a precisão dos sete algoritmos e construirá um nugget de modelo único contendo os três mais precisos.

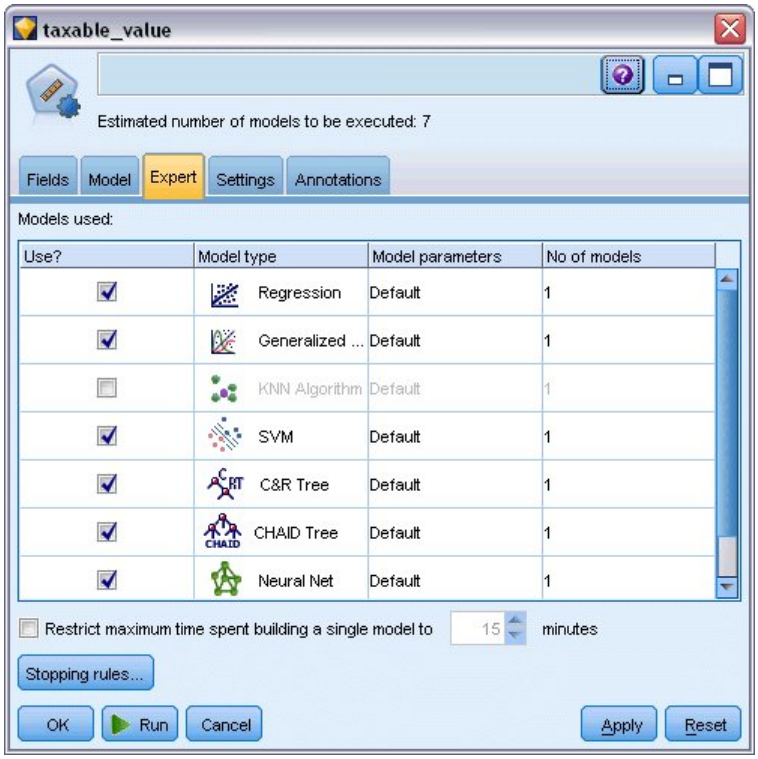


Figura 46. Guia Numérico de nó Numérico Expert

6. Na guia Configurações, deixe as configurações padrão no lugar. Como esta é uma meta contínua, a pontuação da combinação é gerada pela média das pontuações dos modelos individuais.

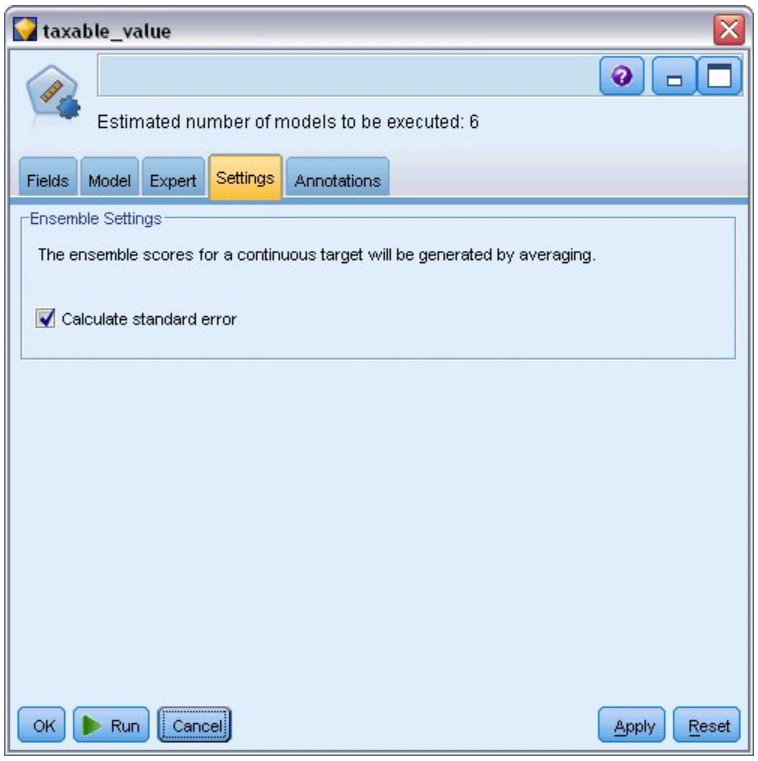


Figura 47. Guia Configurações do nó Numérico automático

Comparando os modelos

1. Clique no botão Executar.

O nugget modelo é construído e colocado na tela, e também na paleta de Models no canto superior direito da janela. Você pode navegar no nugget, ou salvar ou implementá-lo de uma série de outras maneiras.

Abra o nugget modelo; ele lista detalhes sobre cada um dos modelos criados durante a execução. (Em uma situação real, na qual centenas de modelos são estimados em um grande dataset, isso pode levar muitas horas.) Consulte Figura 42 na página 43.

Se você deseja explorar mais adiante qualquer um dos modelos individuais, você pode clicar duas vezes em um ícone de nugget de modelo na coluna **Modelo** para realizar a pesquisa e navegar nos resultados do modelo individual; a partir daí você pode gerar nós de modelagem, nuggets de modelo ou gráficos de avaliação.

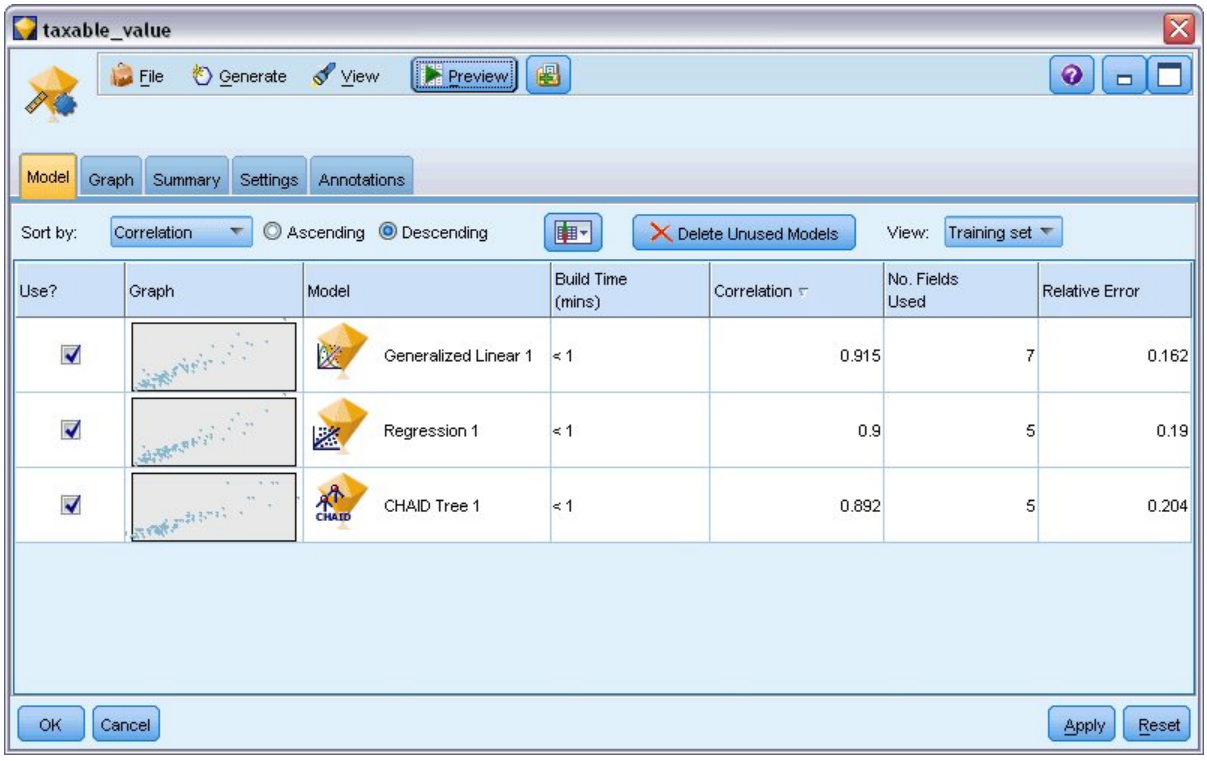


Figura 48. Resultados numéricos automáticos

Por padrão, os modelos são classificados por correlação porque esta foi a medida que você selecionou no nó Numeric automático. Para fins de classificação, utiliza-se o valor absoluto da correlação, com valores mais próximos de 1 indicando uma relação mais forte. O Modelo Linear Generalizado classifica-se melhor nesta medida, mas vários outros são quase tão precisos. O Modelo Linear Generalizado também tem o menor erro relativo.

Você pode classificar em uma coluna diferente clicando no cabeçalho para essa coluna, ou você pode escolher a medida desejada a partir da lista **Classificar por** na barra de ferramentas.

Cada gráfico exibe um enredo de valores observados contra valores previstos para o modelo, proporcionando uma rápida indicação visual da correlação entre eles. Para um bom modelo, os pontos devem se aglomerar ao longo da diagonal, o que é verdadeiro para todos os modelos neste exemplo.

Na coluna **Gráfico** , é possível clicar duas vezes em uma miniatura para gerar um gráfico de tamanho médio.

Com base nesses resultados, você decide usar todos os três desses modelos mais precisos. Ao combinar predições a partir de vários modelos, as limitações em modelos individuais podem ser evitadas, resultando em uma precisão geral mais alta.

No **Use?** coluna, certise-se de que todos os três modelos são selecionados.

Anexar um nó de Análise (paleta de saída) após o nugget modelo. Clique com o botão direito do mouse sobre o nó da Análise e escolha **Executar** para executar o fluxo.

A pontuação média gerada pelo modelo de combinação é incluída em um campo denominado *$XR-taxable_value*, com uma correlação de 0.922, que é maior que aqueles dos três modelos individuais. As pontuações da combinação também mostram um erro absoluto médio baixo e podem ter um desempenho melhor do que qualquer um dos modelos individuais quando aplicado a outros conjuntos de dados.

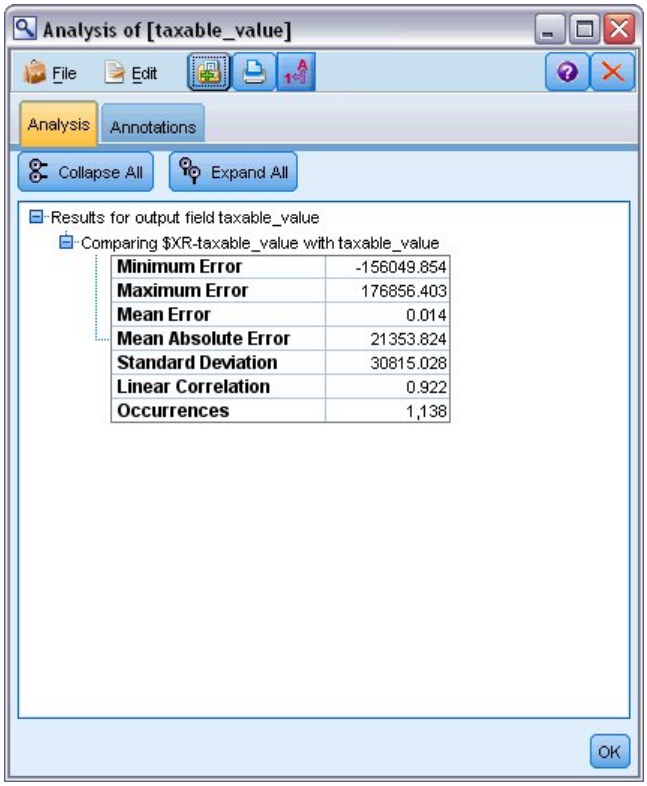


Figura 49. Fluxo De Amostra Numérico Automático

Resumo

Para resumir, você usou o nó Auto Numeric para comparar uma série de modelos diferentes, selecionou os três modelos mais precisos e os adicionou ao fluxo dentro de uma nugget de modelo Auto Numeric ensembled.

- Com base na precisão geral, os modelos Generalized Linear, Regression e CHAID apresentaram melhor desempenho nos dados de treinamento.
- O modelo combinado mostrou um desempenho melhor do que dois dos modelos individuais e pode ter um desempenho melhor quando aplicado a outros conjuntos de dados. Se seu objetivo é automatizar o processo tanto quanto possível, esta abordagem permite que você obtenha um modelo robusto na maioria das circunstâncias, sem ter que se aprofundar nas especificações de qualquer modelo.

Capítulo 6. Preparação De Dados Automatizados (ADP)

Preparar dados para análise é uma das etapas mais importantes em qualquer projeto de mineração de dados - e tradicionalmente, uma das mais demoradas. O Nó de Preparação de Dados Automatizado (ADP) trata da tarefa para você, analisando seus dados e identificando correções, rastreando campos que são problemáticos ou não prováveis de serem úteis, derivando novos atributos quando apropriado, e melhorando o desempenho através de técnicas de rastreamento inteligentes. Você pode usar o nó de forma totalmente automatizada, permitindo que o nó escolha e aplique correções ou você possa visualizar as alterações antes que elas sejam feitas e aceitá-las ou rejeitá-las conforme desejado.

O uso do nó ADP possibilita que você faça seus dados prontos para mineração de dados de forma rápida e fácil, sem precisar ter conhecimento prévio dos conceitos estatísticos envolvidos. Se você executar o nó com as configurações padrão, os modelos tenderão a ser construídos e pontuados mais rapidamente.

Esse exemplo usa o fluxo denominado *ADP_basic_demo.str*, que faz referência ao arquivo de dados denominado *telco.sav* para demonstrar a precisão aumentada que pode ser localizada usando as configurações do nó ADP padrão ao construir modelos Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *ADP_basic_demo.str* está no diretório *stream* .

Construindo o Fluxo

1. Para construir o fluxo, inclua um nó de origem do File File apontando para `telco.sav` localizado no diretório `Demos` da sua instalação IBM SPSS Modelador .

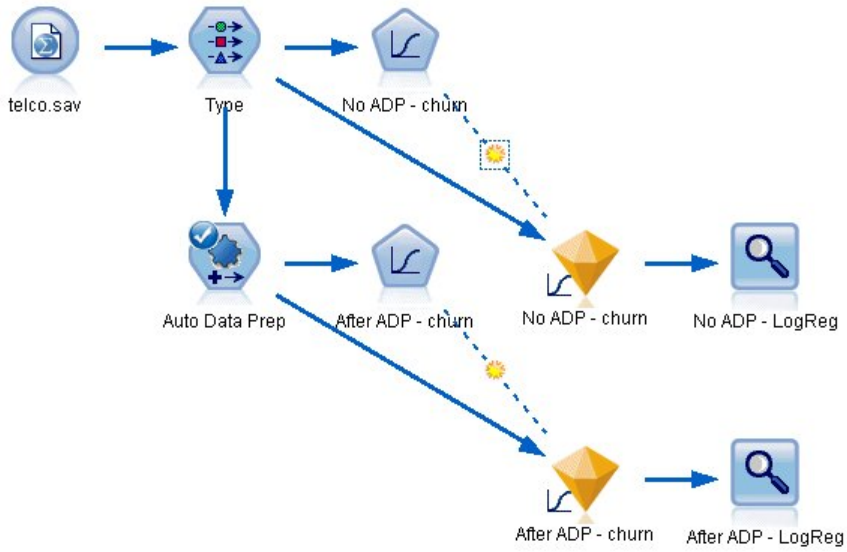


Figura 50. Construindo o Fluxo

2. Conecte um nó Type ao nó de origem, configure o nível de medição para o campo *churn* para **Flag**, e configure a função para **Target**. Todos os outros campos devem ter seu papel configurado como **Entrada**.

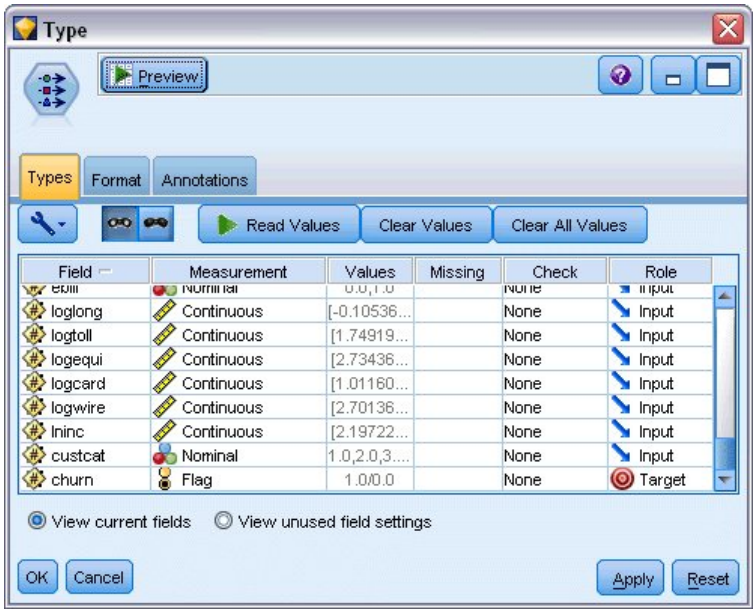


Figura 51. Selecionando o destino

3. Conecte um nó Logística ao nó Tipo.
4. No nó Logístico, clique na guia Modelo e selecione o procedimento **Binomial** . No campo *Nome do modelo* , selecione **Customizado** e insira `No ADP - churn`.

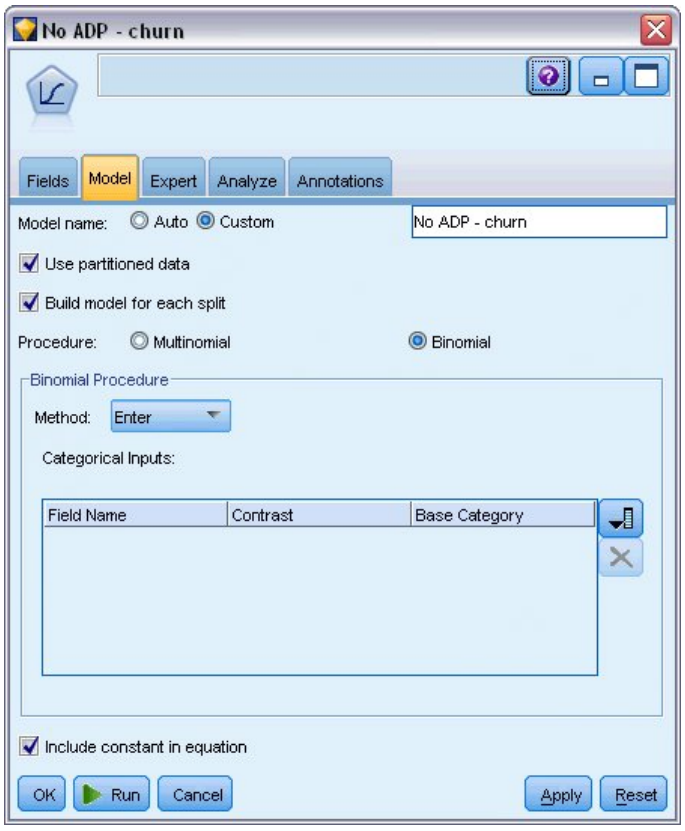


Figura 52. Escolhendo opções de modelo

5. Conecte um nó ADP ao nó Type. Na guia Objetivos, Deixe as configurações padrão em local para analisar e preparar seus dados, equilibrando tanto a velocidade quanto a precisão.
6. Na parte superior da guia Objetivos, Clique em **Analisar Dados** para analisar e processar seus dados.

Outras opções no nó ADP permitem que você especifique que deseja se concentrar mais na precisão, mais na velocidade de processamento ou para ajustar bem muitas das etapas de processamento de preparação de dados.

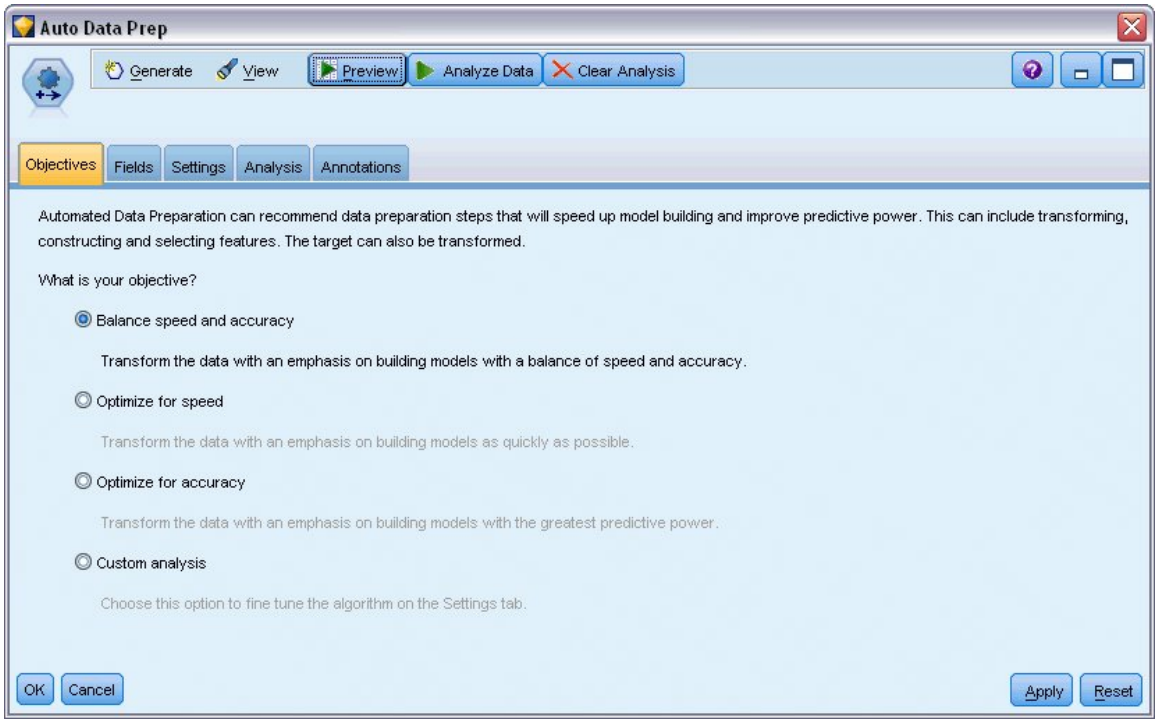


Figura 53. Objetivos padrão de ADP

Os resultados do processamento de dados são exibidos na guia Análise. O **Resumo de Processamento de Campo** mostra que dos 41 recursos de dados trazidos para o nódulo ADP, 19 foram transformados em processamento de ajuda, sendo que 3 foram descartados como não utilizados.

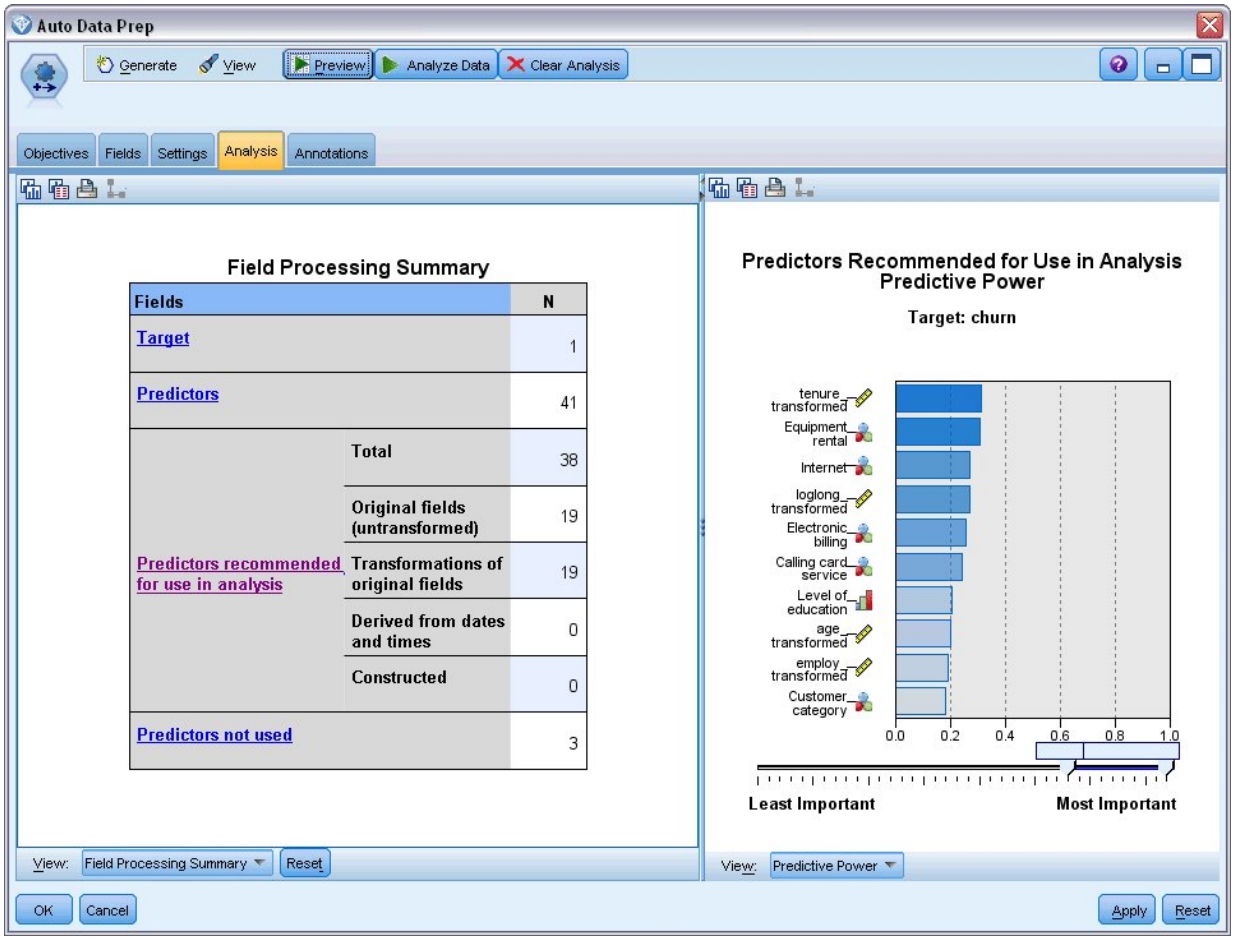


Figura 54. Resumo do processamento de dados

7. Conecte um nó Logístico ao nó ADP.
8. No nó Logístico, clique na guia Modelo e selecione o procedimento **Binomial** . No campo *Nome da Modelagem* selecione **Customizado** e insira `After ADP - churn`..

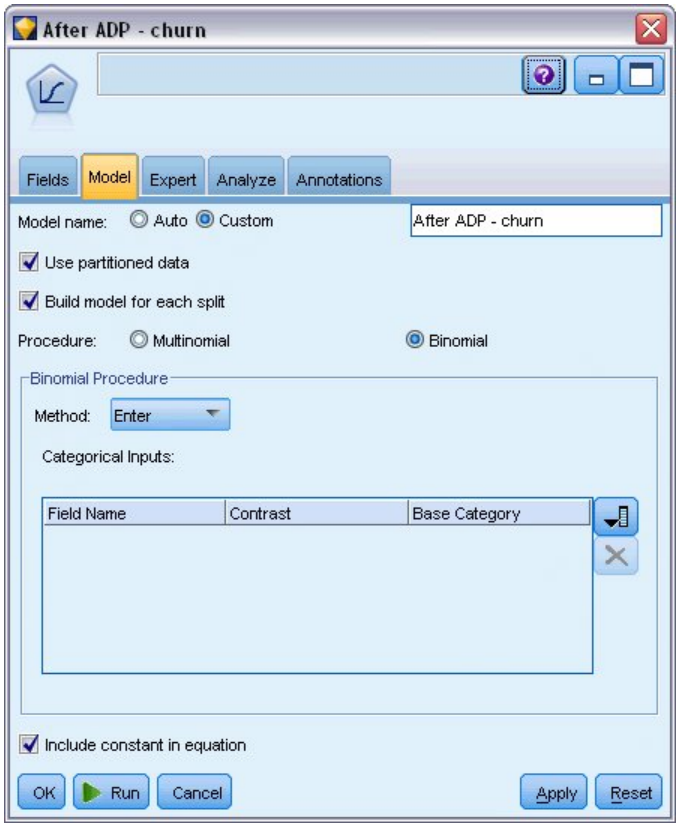


Figura 55. Escolhendo opções de modelo

Comparando Precisão Do Modelo

1. Execute ambos os nós Logísticos para criar os nuggets do modelo, que são adicionados ao fluxo e à paleta de Modelos no canto superior direito.

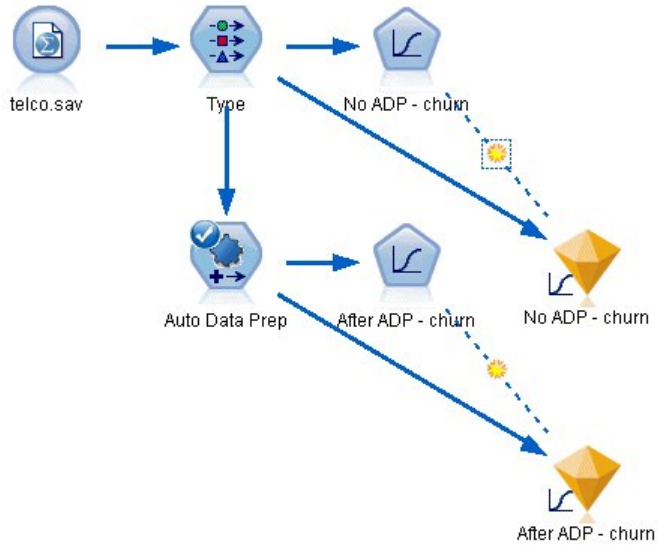


Figura 56. Anexando os nuggets do modelo

2. Conecte nós de Análise aos nuggets de modelo e execute os nós de Análise usando suas configurações padrão.

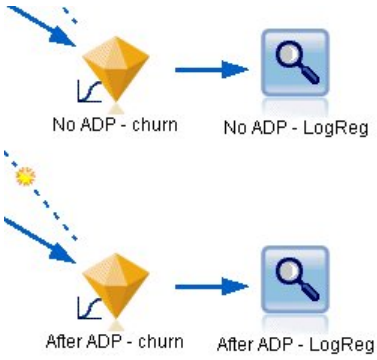


Figura 57. Anexando os nós de análise

A Análise do modelo não derivado de ADP mostra que apenas executar os dados por meio do nó de Regressão Logística com suas configurações padrão fornece um modelo com baixa precisão-apenas 10.6%.

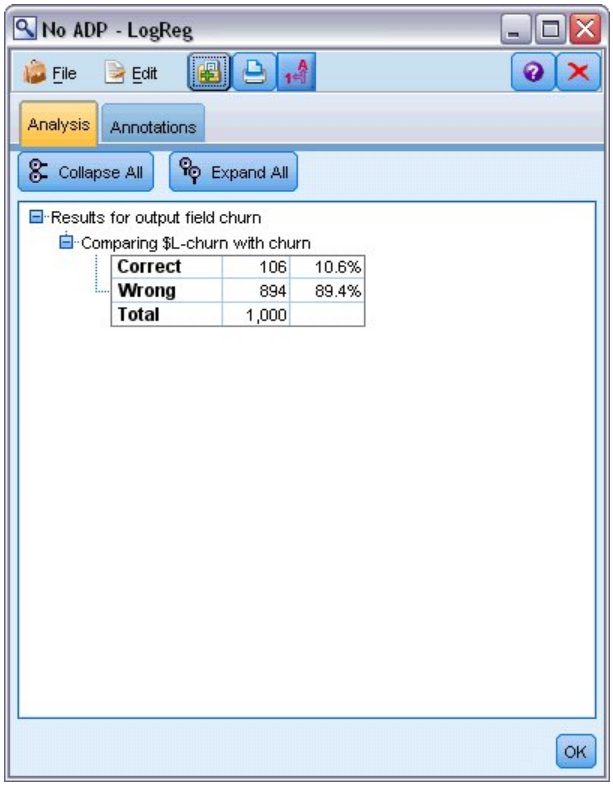


Figura 58. Resultados do modelo não derivado da preparação automática de dados

A Análise do modelo derivado do ADP mostra que executando os dados por meio das configurações padrão do ADP, você construiu um modelo muito mais preciso que está 78.8% correto.

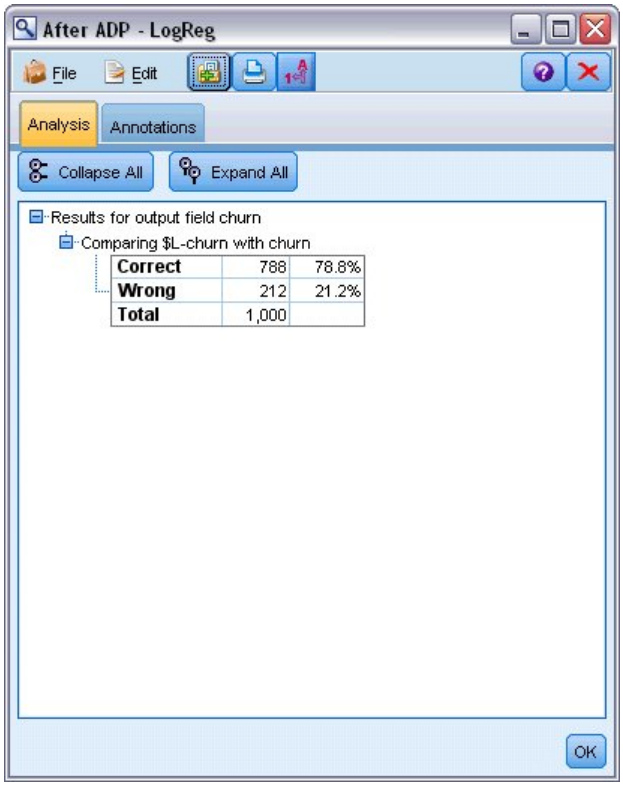


Figura 59. Resultados do modelo derivado da preparação automática de dados

Em resumo, ao apenas executar o nó ADP para ajustar fino o processamento de seus dados, você foi capaz de construir um modelo mais preciso com pouca manipulação de dados diretos.

Obviamente, se você está interessado em provar ou desprovar uma determinada teoria, ou deseja construir modelos específicos, você pode achar benéfica trabalhar diretamente com as configurações do modelo; no entanto, para aqueles com um tempo reduzido de tempo, ou com uma grande quantidade de dados para preparar, o nó ADP pode dar uma vantagem.

Explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no *IBM SPSS Modelador Algorithms Guide*, disponível a partir do diretório *\Documentação* do disco de instalação.

Observe que os resultados neste exemplo são baseados apenas nos dados de treinamento. Para avaliar o quão bem os modelos generalizam para outros dados no mundo real, você usaria um nó de partição para conter um subconjunto de registros para fins de teste e validação.

Capítulo 7. Preparando Dados para Análise (Data Audit)

O nó de auditoria de Dados fornece um primeiro olhar abrangente sobre os dados que você traz para o IBM SPSS Modelador. Frequentemente usado durante a exploração inicial de dados, o relatório de auditoria de dados mostra estatísticas de resumo, bem como histogramas e gráficos de distribuição para cada campo de dados, e permite especificar tratamentos para valores ausentes, outliers e valores extremas.

Este exemplo usa o fluxo denominado *telco_dataaudit.str*, que faz referência ao arquivo de dados denominado *telco.sav* Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *telco_dataaudit.str* está no diretório *streams* .

Construindo o Fluxo

1. Para construir o fluxo, inclua um nó de origem do Arquivo de Estatísticas apontando para *telco.sav* localizado no diretório *Demos* de sua instalação do IBM SPSS Modelador .

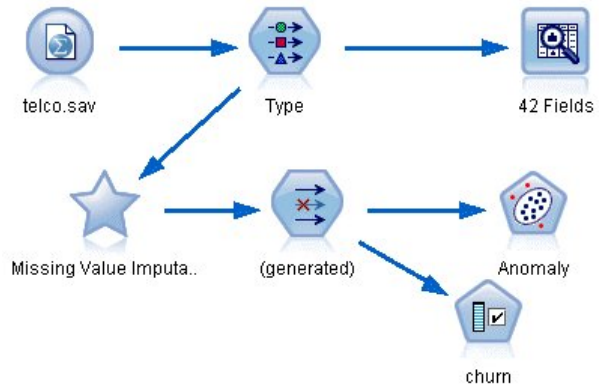


Figura 60. Construindo o Fluxo

2. Adiciona um nó Type para definir campos, e especifique *churn* como o campo de destino (Role = **Target**). Função deve ser definida como **Entrada** para todos os outros campos para que este seja o único alvo.

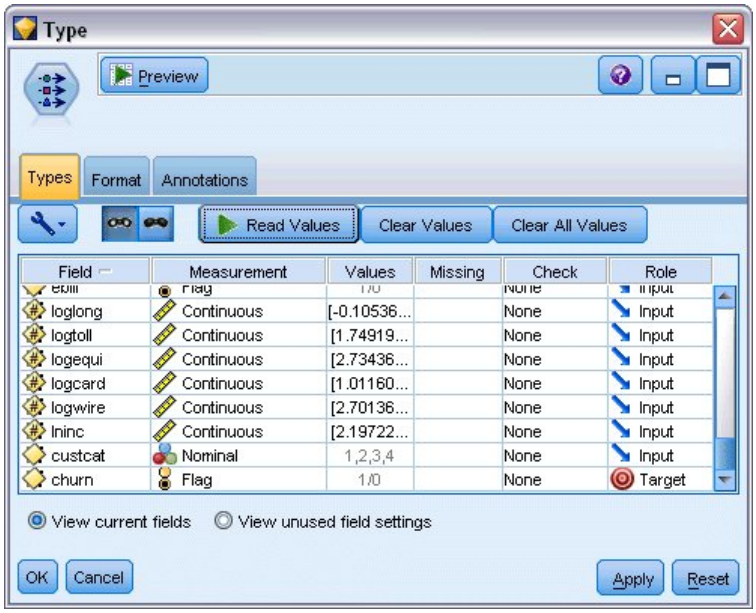


Figura 61. Configurando o destino

3. Confirme se os níveis de medição de campo estão definidos corretamente. Por exemplo, a maioria dos campos com valores 0 e 1 pode ser considerada como sinalizadores, mas certos campos, como o gênero, são visualizados com mais precisão como um campo nominal com dois valores.

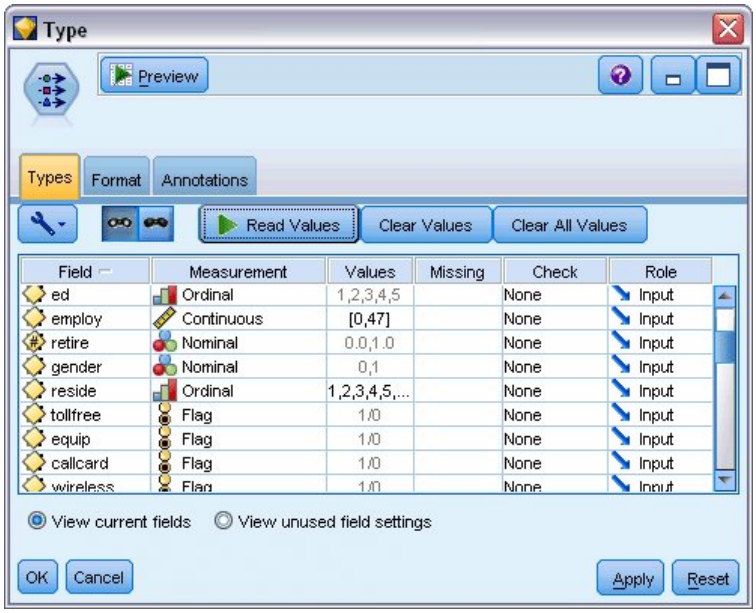


Figura 62. Configurando níveis de medição

Dica: Para alterar propriedades para vários campos com valores semelhantes (tais como 0/ 1), clique no cabeçalho da coluna *Valores* para classificar campos por essa coluna, e use a tecla Shift para selecionar todos os campos que deseja alterar. Em seguida, é possível clicar com o botão direito do mouse sobre a seleção para alterar o nível de medição ou outros atributos para todos os campos selecionados.

4. Anexar um nó de Auditoria De Dados ao fluxo. Na aba Configurações, deixe as configurações padrão no lugar para incluir todos os campos do relatório. Já que *churn* é o único campo de destino definido no nó Type, ele será automaticamente usado como uma sobreposição.

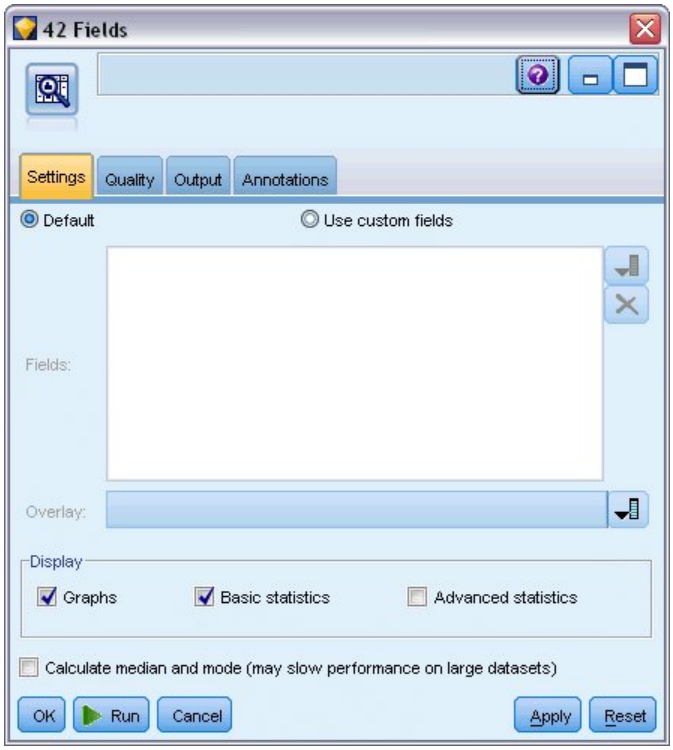


Figura 63. Nó de auditoria de dados, guia Configurações

Na guia Qualidade, deixe as configurações padrão para detecção de valores ausentes, outliers e valores extremas no lugar e clique em **Executar**.

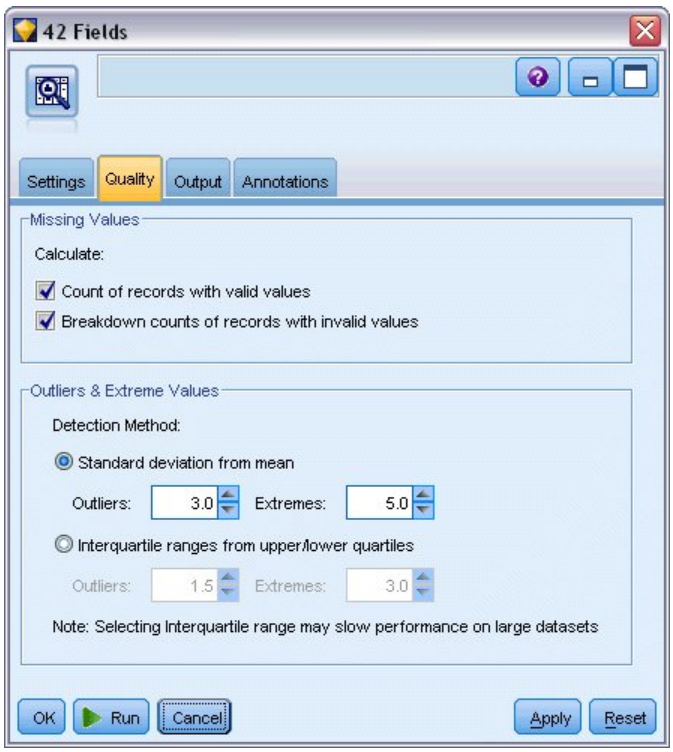


Figura 64. Nó de auditoria de dados, guia de qualidade

Estatísticas de Navegação e Gráficos

O navegador Data Audit é exibido, com gráficos em miniatura e estatística descritiva para cada campo.

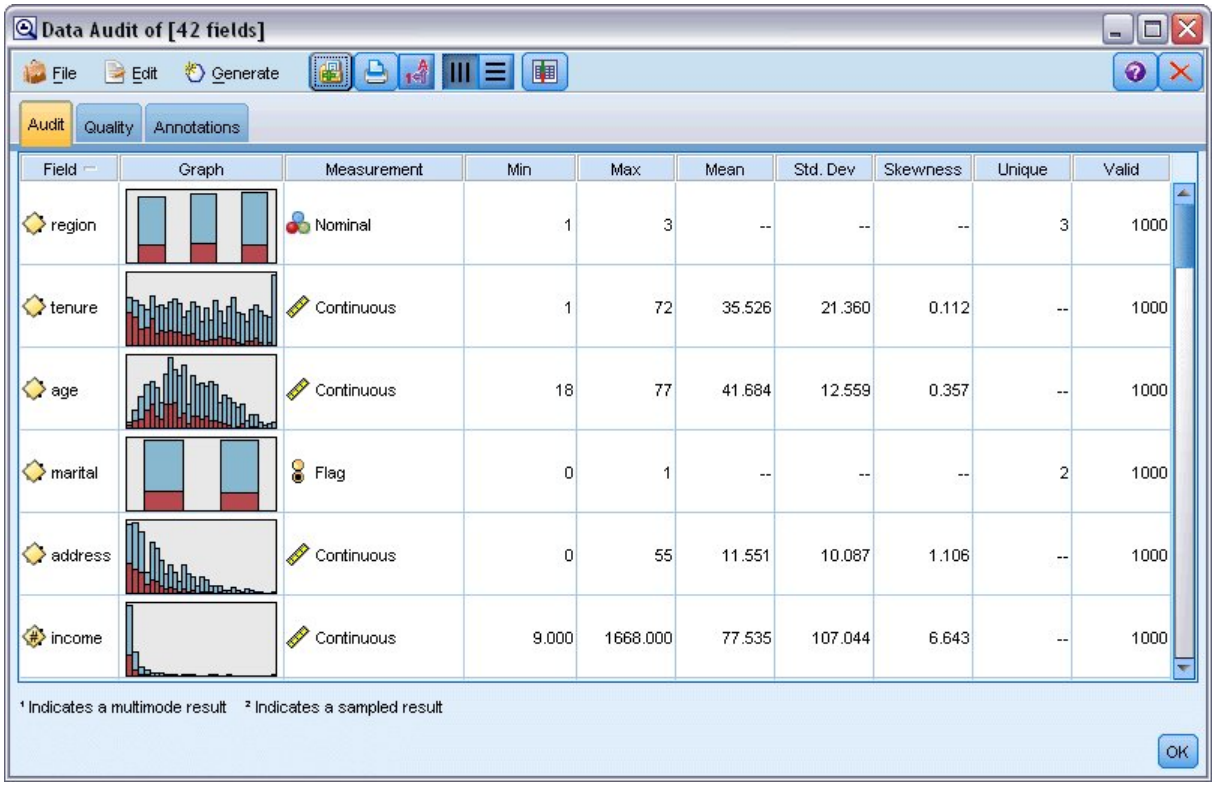


Figura 65. navegador de Auditoria de Dados

Use a barra de ferramentas para exibir etiquetas de campo e de valor e para alternar o alinhamento de gráficos da horizontal para vertical (apenas para campos categóricos).

1. Você também pode usar a barra de ferramentas ou o menu Editar para escolher as estatísticas a serem exibidas.

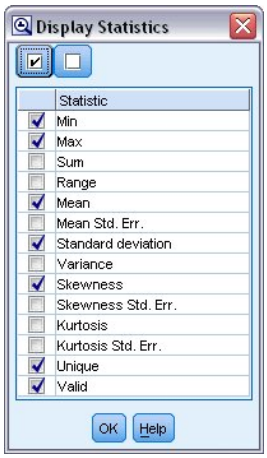


Figura 66. Exibir Estatísticas

Clique duas vezes em qualquer gráfico de miniatura no relatório de auditoria para visualizar uma versão de tamanho completo desse gráfico. Como *churn* é o único campo de destino no fluxo, ele é usado automaticamente como uma sobreposição. Você pode alternar a exibição de etiquetas de campo e de valor usando a barra de ferramentas da janela do gráfico, ou clicar no botão Editar modo para customizar ainda mais o gráfico.

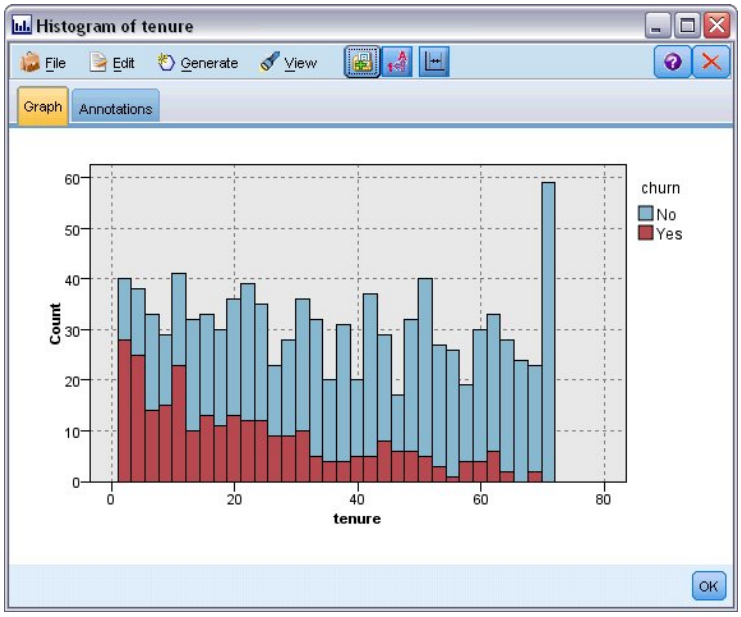


Figura 67. Histograma de tenure

Alternativamente, você pode selecionar uma ou mais miniaturas e gerar um nó Gráfico para cada um. Os nós gerados são colocados na tela do fluxo e podem ser adicionados ao fluxo para recriar esse gráfico específico.

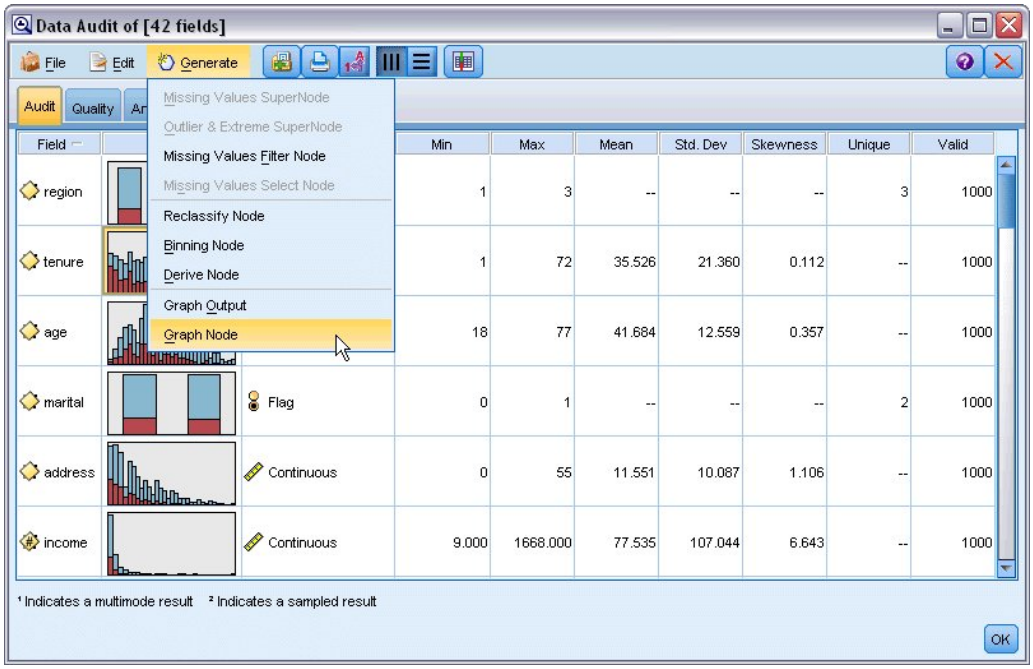


Figura 68. Gerando um Nó Gráfico

Cuidando de Outliers e Valores Ausentes

A guia Qualidade no relatório de auditoria exibe informações sobre outliers, extremos e valores ausentes.

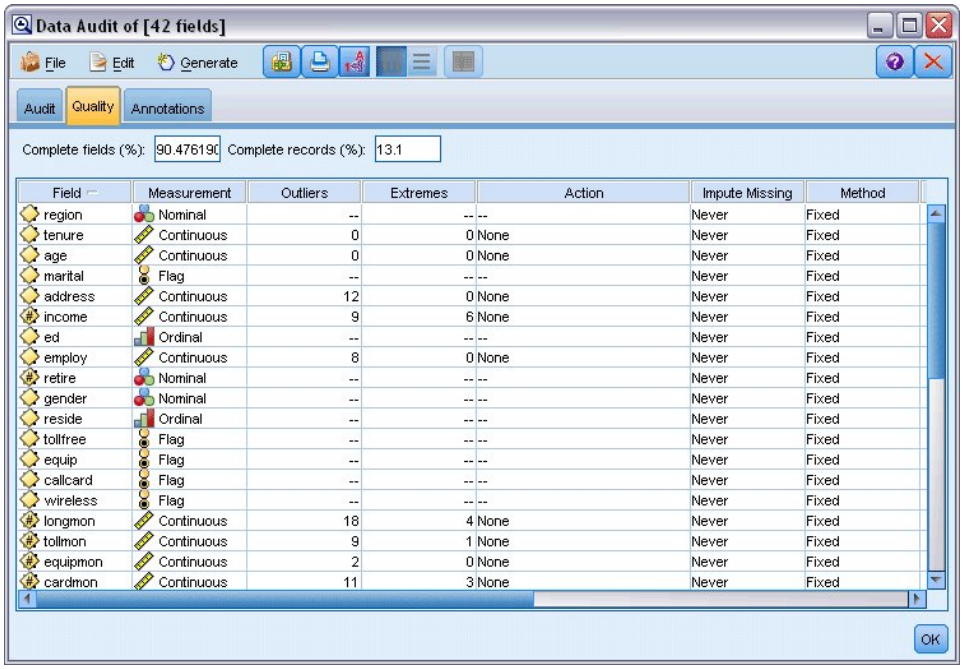


Figura 69. Navegador de auditoria de dados, guia de qualidade

Também é possível especificar métodos para manipular esses valores e gerar SuperNodes para aplicar automaticamente as transformações.. Por exemplo você pode selecionar um ou mais campos e optar por imputar ou substituir valores ausentes para esses campos usando uma série de métodos, incluindo o algoritmo C & RT.

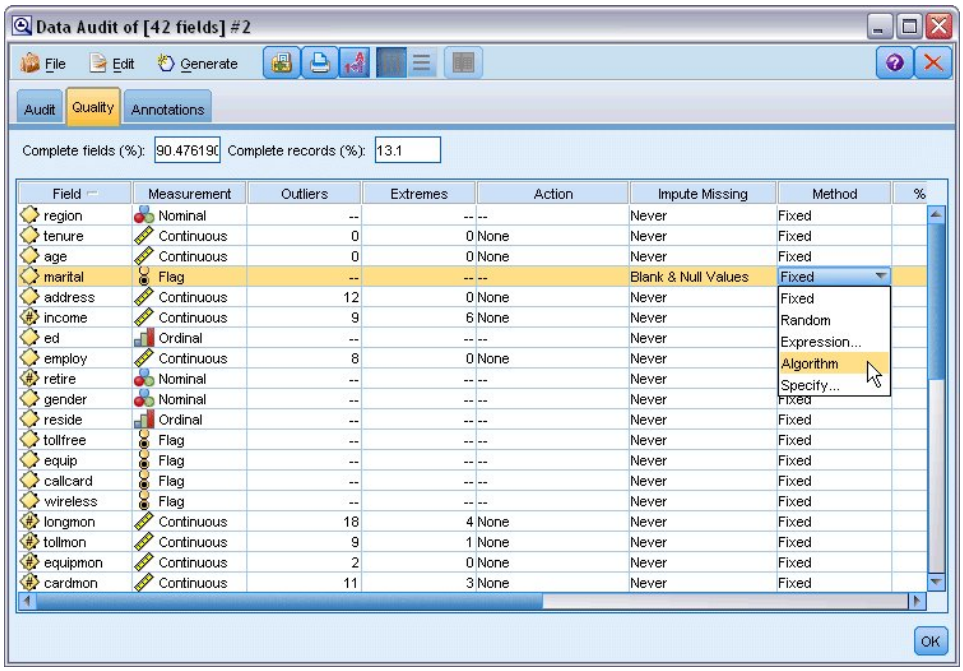


Figura 70. Escolhendo um método de imputação

Depois de especificar um método de imputação para um ou mais campos, para gerar um " SuperNode, de valores ausentes, selecione nos menus:

Gerar > **Valores Omissos SuperNode**

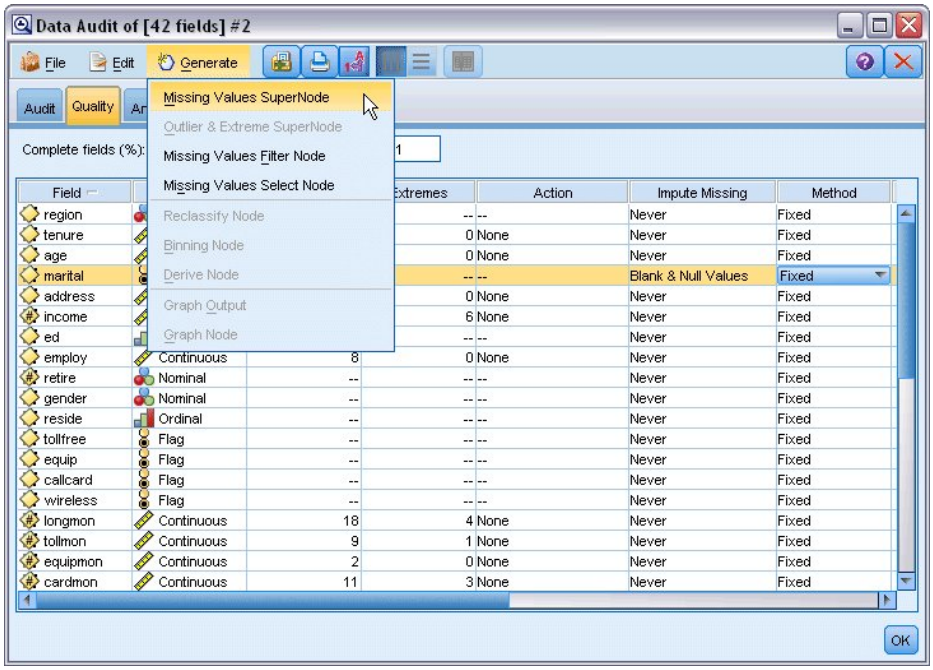


Figura 71. Gerando o SuperNode

O SuperNode gerado é incluído na tela de fluxo, na qual é possível anexá-lo ao fluxo para aplicar as transformações

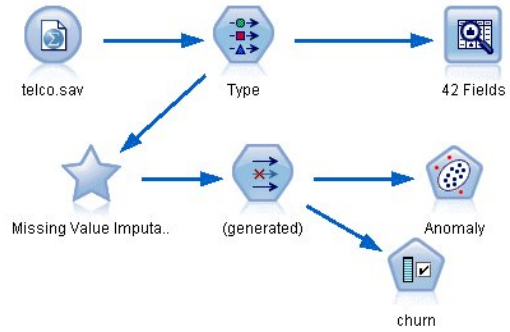


Figura 72. Fluxo com valores omissos SuperNode

O SuperNode realmente contém uma série de nós que executam as transformações solicitadas. Para entender como ele funciona, é possível editar o SuperNode e clicar em **Zoom In**.

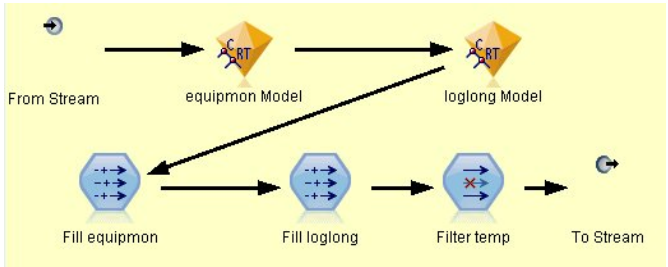


Figura 73. Aumentando o zoom no SuperNode

Para cada campo imputado usando o método do algoritmo, por exemplo, haverá um modelo C & RT separado, juntamente com um nó Filler que substitui espaços em branco e nulos com o valor previsto pelo modelo. É possível incluir, editar ou remover nós específicos no SuperNode para customizar mais o comportamento.

Alternativamente, você pode gerar um nó Select ou Filter para remover campos ou registros com valores ausentes. Por exemplo, você pode filtrar quaisquer campos com uma porcentagem de qualidade abaixo de um limite especificado.

Figura 74. Gerando um Nó Filtro

Outliers e valores extremas podem ser tratados de maneira semelhante. Especifique a ação que deseja tomar para cada campo-coerção, descarte ou anulação-e gere um SuperNode para aplicar as transformações.

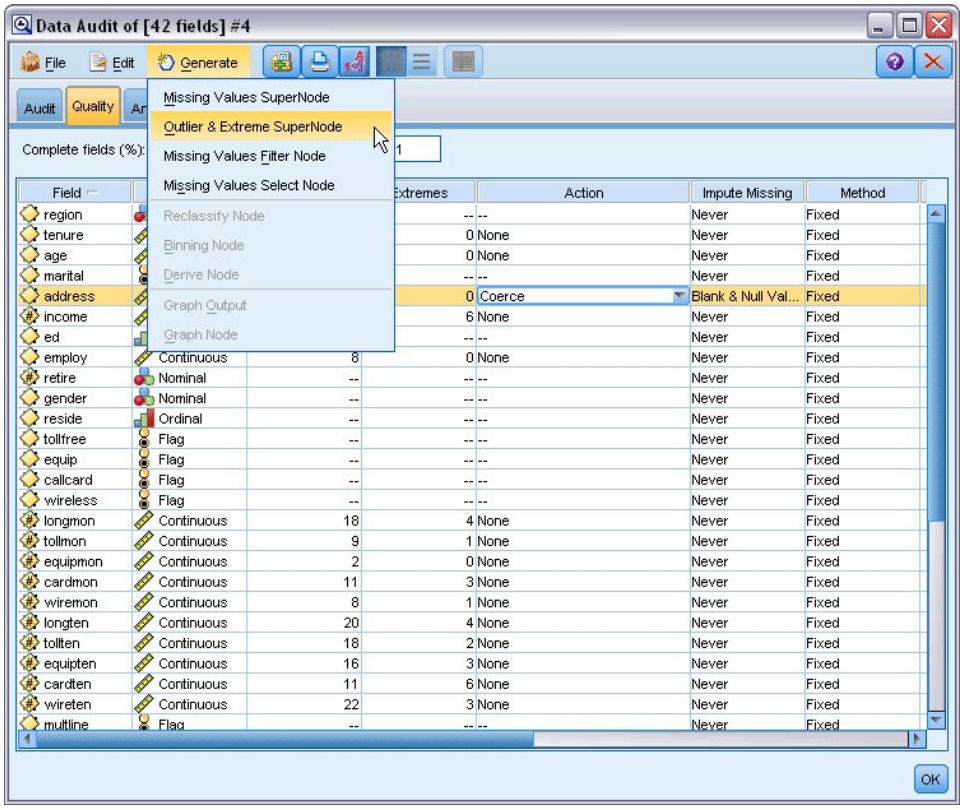


Figura 75. Gerando um Nó Filtro

Após concluir a auditoria e adicionar os nós gerados ao fluxo, você pode prosseguir com a sua análise. Opcionalmente, você pode desejar mais tela seus dados usando Detecção Anomalia, Seleção de Recurso ou uma série de outros métodos.

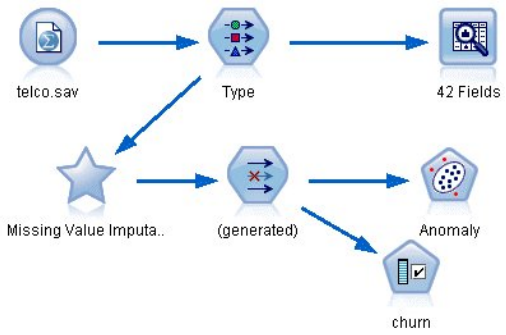


Figura 76. Fluxo com valores omissos SuperNode

Capítulo 8. Tratamentos De Drogas (Exploratório Graphs/C5.0)

Para esta seção, imagine que você é um pesquisador médico compilando dados para um estudo. Você coletou dados sobre um conjunto de pacientes, todos os quais sofreram com a mesma doença. Durante o curso do tratamento, cada paciente respondeu a um de cinco medicamentos. Parte do seu trabalho é usar a mineração de dados para descobrir qual medicamento pode ser apropriado para um futuro paciente com a mesma doença.

Este exemplo usa o fluxo denominado *druglearn.str*, que faz referência ao arquivo de dados denominado *DRUG1n*. Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *druglearn.str* está no diretório *streams*

Os campos de dados utilizados na demo são:

Campo de dados	Descrição
Idade	(Número)
Sexo	*M* ou *F*
BP	Pressão arterial: *HIGH*, *NORMAL* ou *LOW*
Colesterol	Colesterol no sangue: *NORMAL* or *HIGH*
Na	Concentração de sódio no sangue
k	Concentração de potássio no sangue
Medicamento	Medicamento prescrito ao qual um paciente respondeu

Leitura de dados de texto

Figura 77. Adicionando um Nó de Arquivo Variável

Você pode ler em dados de texto delimitados usando um **Nó de Arquivo Variável**. Você pode adicionar um nó do File Variable a partir das paletas -- ou clique na aba **Fontes** para localizar o nó ou usar a guia **Favoritos** , que inclui este nó por padrão. Em seguida, dê um duplo clique no nó recém-colocado para abrir sua caixa de diálogo.

Clique no botão apenas para a direita da caixa de arquivo marcada com um ellipsis (...) para navegar até o diretório no qual IBM SPSS Modelador está instalado em seu sistema. Abra o diretório *Demos* e selecione o arquivo chamado *DRUG1n*.

Garantindo que **Leia nomes de campos do arquivo** esteja selecionado, observe os campos e valores que acabaram de ser carregados na caixa de diálogo.

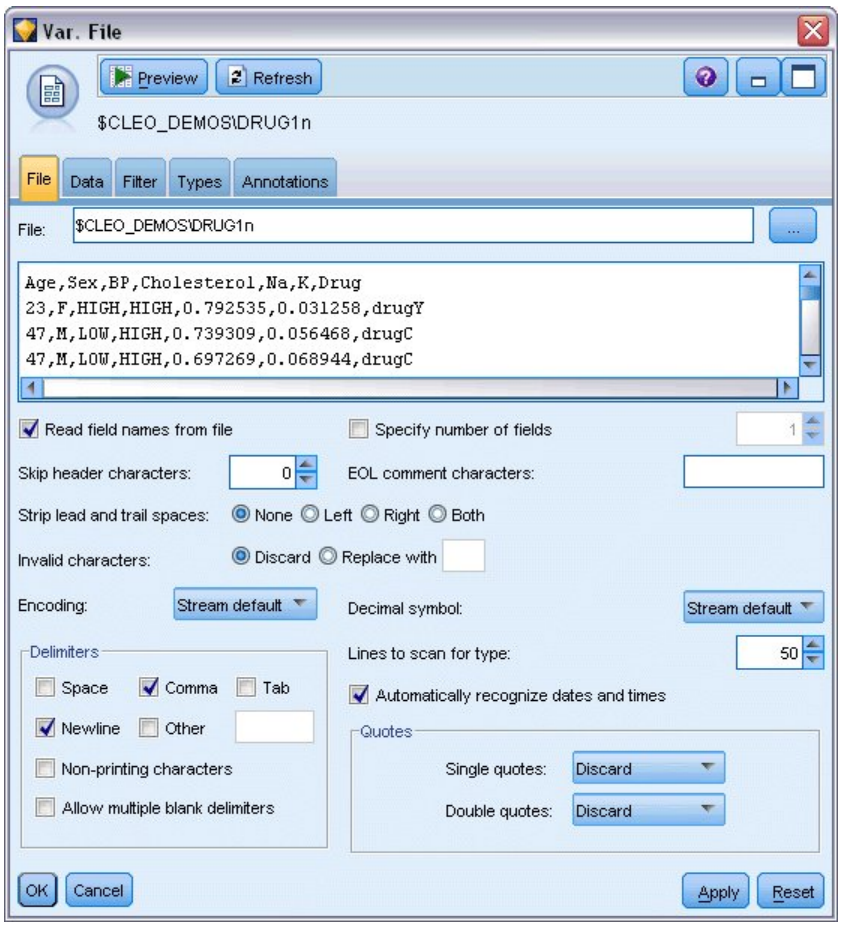


Figura 78. Caixa de diálogo de arquivo variável

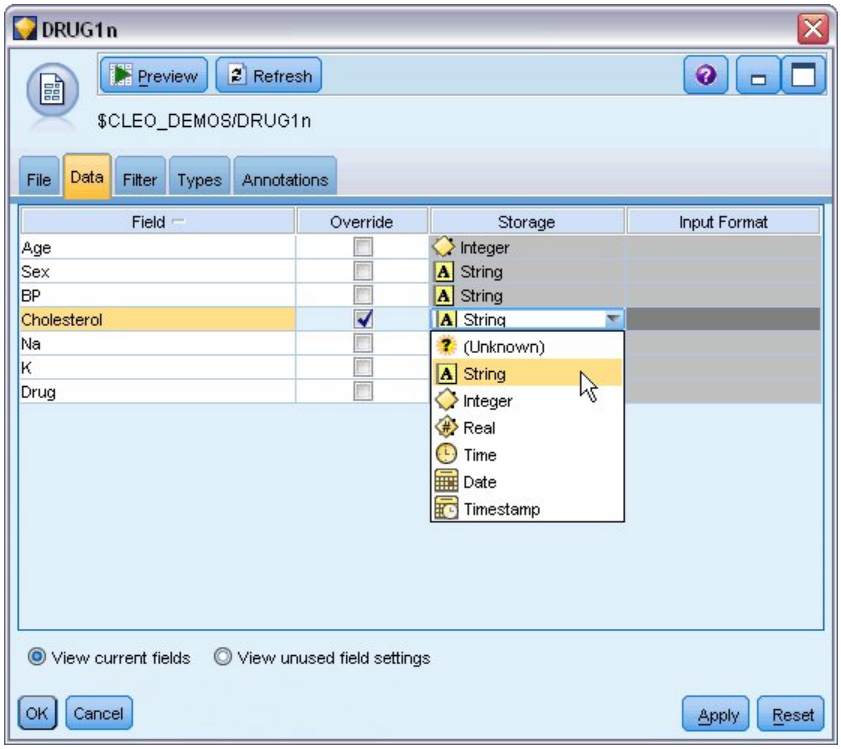


Figura 79. Como alterar o tipo de armazenamento para um campo

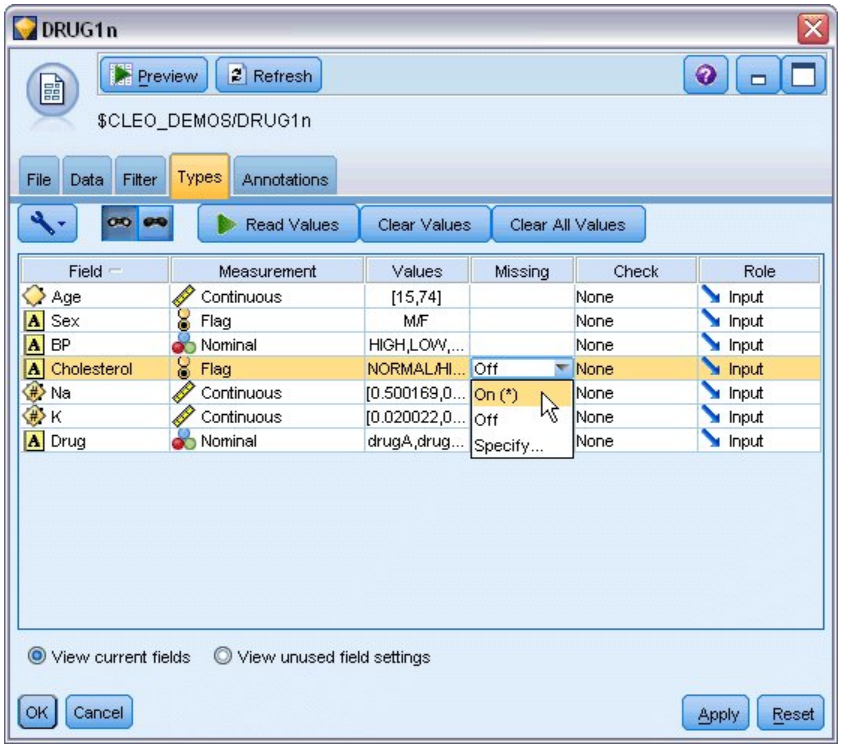


Figura 80. Seleção de opções de Valor na guia Tipos

Clique na aba **Dados** para substituir e alterar **Armazenamento** para um campo. Note que o armazenamento é diferente de **Measurement**, ou seja, o nível de medição (ou tipo de uso) do campo de dados. A aba **Tipos** ajuda você a saber mais sobre o tipo de campos em seus dados. Você também pode escolher **Valores de leitura** para visualizar os valores reais para cada campo com base nas seleções que você faz da coluna *Valores* . Este processo é conhecido como **instanciation**.

Incluindo uma tabela

Agora que você carregou o arquivo de dados, você pode querer olhar para os valores para alguns dos registros. Uma maneira de fazer isso é construindo um fluxo que inclui um nó da Tabela. Para colocar um nó da Tabela no fluxo, clique duas vezes sobre o ícone na paleta ou arraste e solte-o na tela.

Figura 81. Nó da tabela conectado à fonte de dados

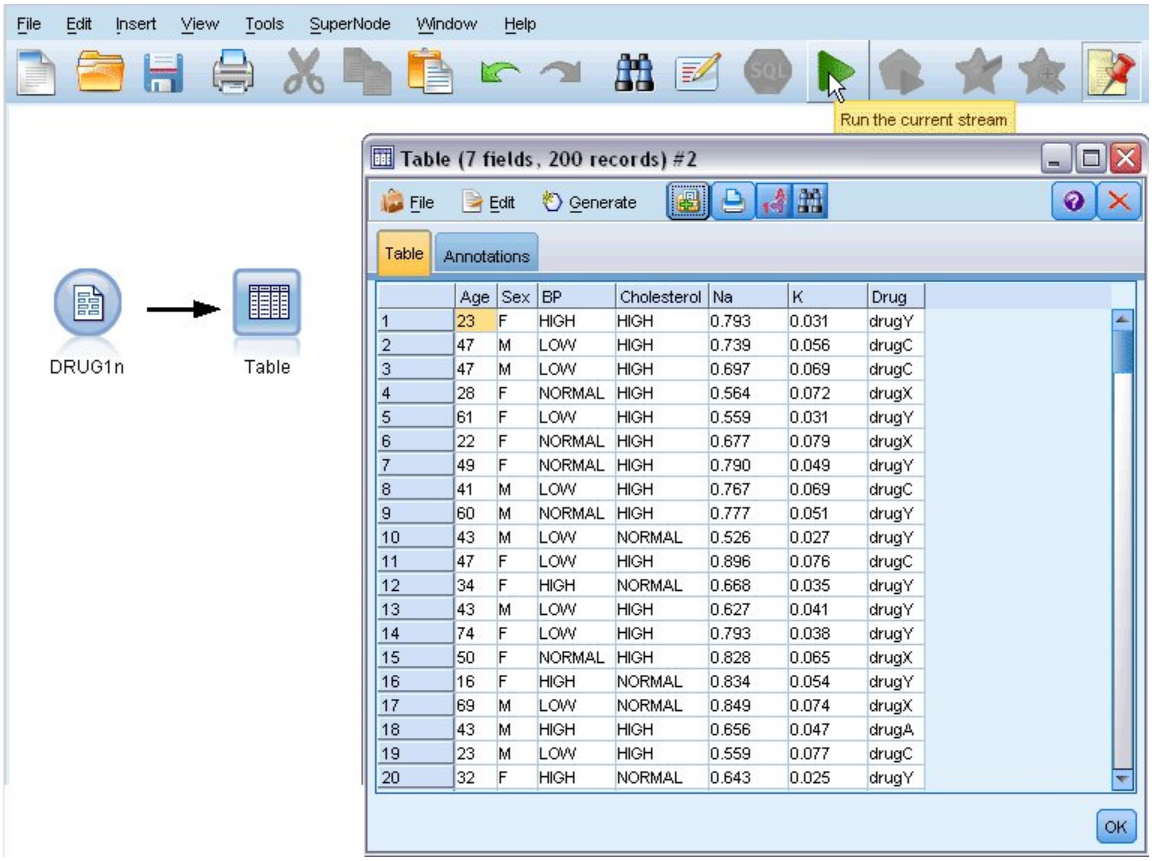


Figura 82. Executando um Fluxo a partir da barra de ferramentas

O duplo clique de um nó da paleta irá conectá-lo automaticamente ao nó selecionado na tela do fluxo. Alternativamente, se os nós já não estiverem conectados, você poderá usar o seu botão do meio do mouse para conectar o nó Fonte ao nó da Tabela. Para simular um botão do meio do mouse, mantenha pressionada a tecla Alt enquanto utiliza o mouse. Para visualizar a tabela, clique no botão de seta verde na barra de ferramentas para executar o stream, ou clique com o botão direito do mouse no nó da Tabela e escolha **Executar**.

Criando um Graph de Distribuição

Durante a mineração de dados, muitas vezes é útil explorar os dados, criando resumos visuais. IBM SPSS Modelador oferece vários tipos diferentes de gráficos para escolher, dependendo do tipo de dados que você deseja resumir. Por exemplo, para descobrir qual proporção dos pacientes respondeu a cada medicamento, use um nó de distribuição.

Inclua um nó de Distribuição no fluxo e conecte-o ao nó Fonte, em seguida, dê um duplo clique no nó para editar opções para exibição.

Selecione *Medicamento* como o campo de destino cuja distribuição você deseja mostrar. Em seguida, clique em **Executar** a partir da caixa de diálogo.

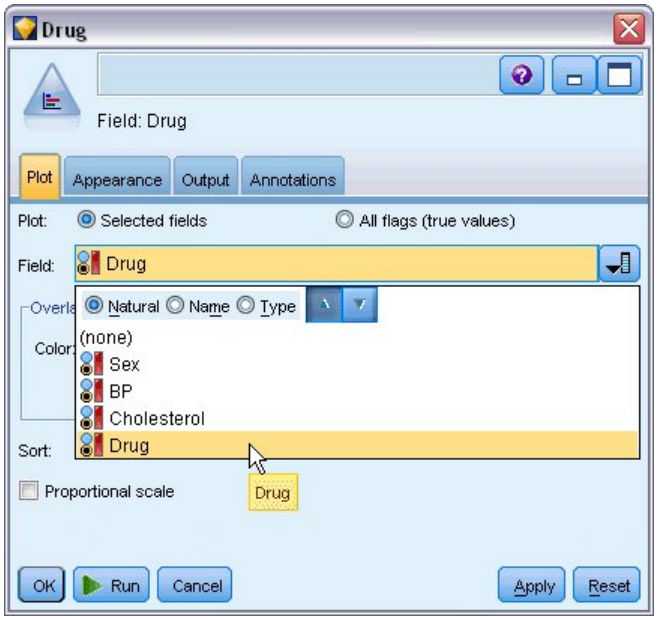


Figura 83. Seleção de droga como campo de destino

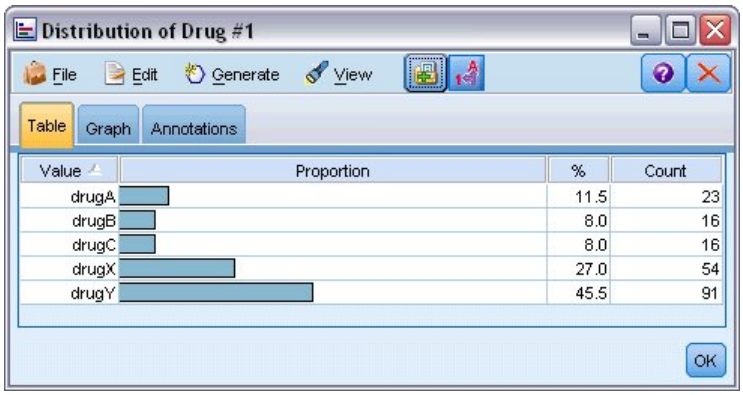


Figura 84. Distribuição de resposta ao tipo de droga

O gráfico resultante ajuda a ver a "forma" dos dados. Isso mostra que os pacientes responderam ao medicamento *Y* com mais frequência e aos medicamentos *B* e *C*com menos frequência.

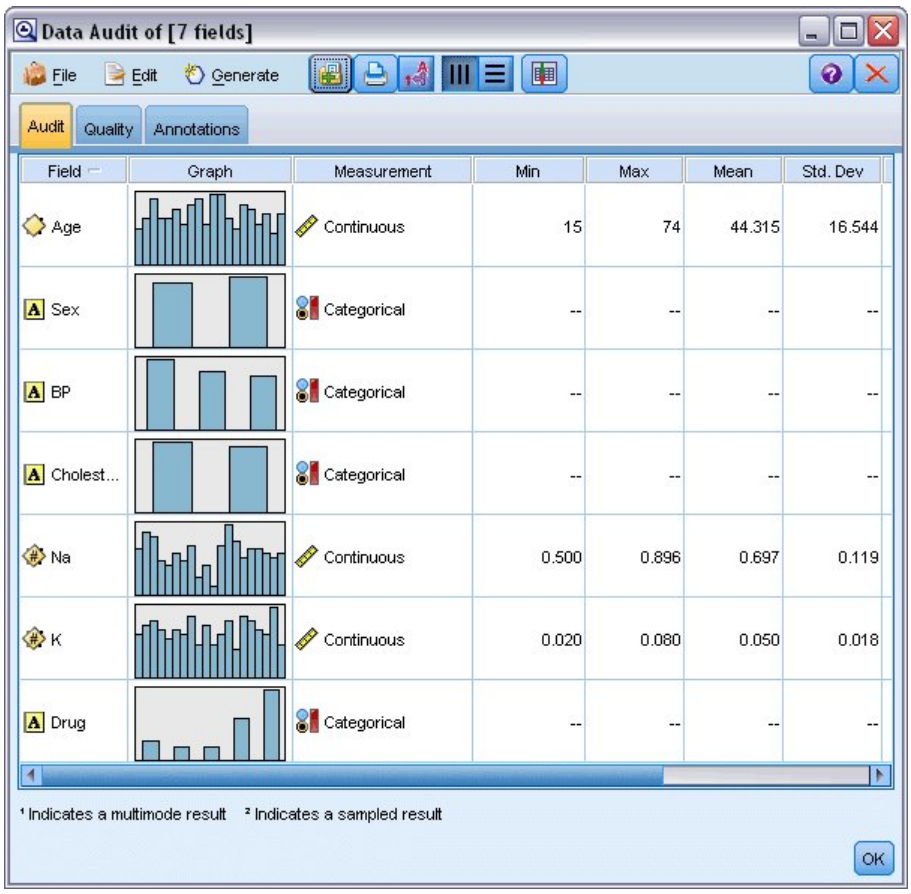


Figura 85. Resultados de uma auditoria de dados

Alternativamente, é possível anexar e executar um nó de auditoria de Dados para uma rápida olhada em distribuições e histogramas para todos os campos de uma só vez. O nó de auditoria de Dados está disponível na aba Saída.

Criando um gráfico de dispersão

Agora, vamos dar uma olhada em quais fatores podem influenciar o *Medicamento*, a variável de destino. Como pesquisador, você sabe que as concentrações de sódio e potássio no sangue são fatores importantes. Uma vez que ambos são valores numéricos, é possível criar um gráfico de dispersão de sódio versus potássio, usando as categorias de medicamentos como uma sobreposição de cores.

Coloque um nó de Plot na área de trabalho e conecte-a ao nó Fonte, e clique duas vezes para editar o nó.

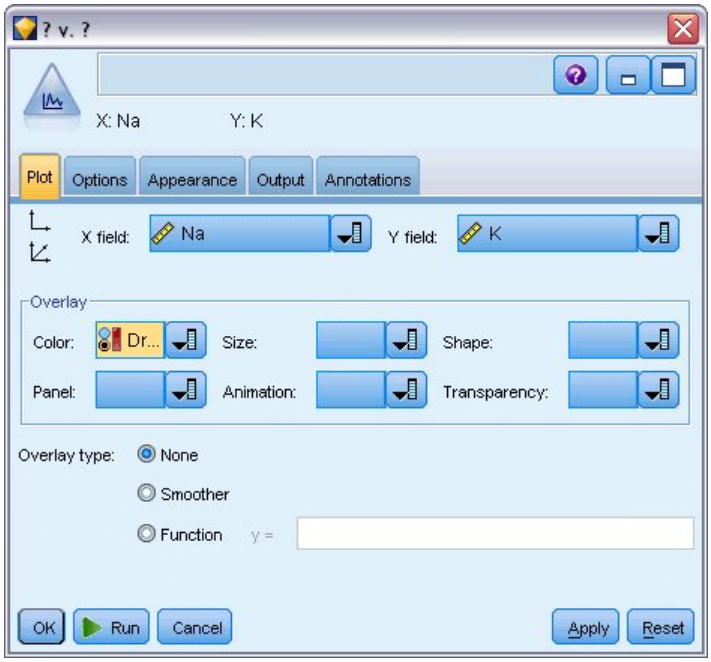


Figura 86. Criando um gráfico de dispersão

Na guia Plot, selecione *Na* como o campo X, *K* como o campo Y, e *Drug* como o campo overlay. Em seguida, clique em **Executar**.

O gráfico mostra claramente um limite acima do qual o medicamento correto é sempre o medicamento *Y* e abaixo do qual o medicamento correto nunca é o medicamento *Y*. Este limiar é uma proporção -- a proporção de sódio (*Na*) para o potássio (*K*).

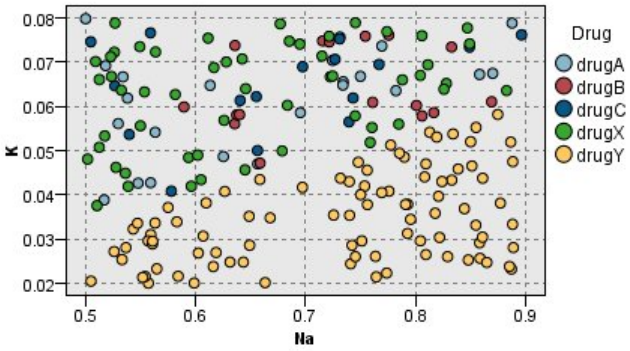


Figura 87. Gráfico de dispersão da distribuição de medicamentos

Criando um Web Graph

Uma vez que muitos dos campos de dados são categóricos, você também pode tentar plotar um gráfico web, que mapeia associações entre diferentes categorias. Inicie conectando um nó da Web ao nó Fonte em sua área de trabalho. Na caixa de diálogo do nó da Web, selecione *BP* (para pressão arterial) e *Drug*. Em seguida, clique em **Executar**.

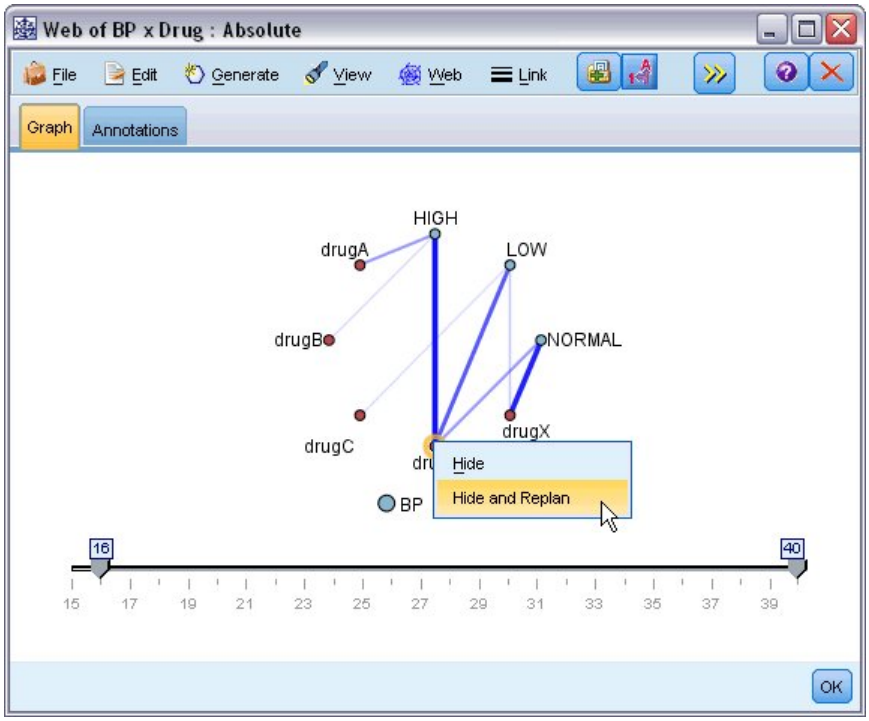


Figura 88. Gráfico da Web de medicamento versus pressão arterial

A partir do gráfico, parece que o medicamento *Y* está associado a todos os três níveis de pressão arterial. Isso não é nenhuma surpresa-você já determinou a situação em que o medicamento *Y* é melhor. Para se concentrar nas outras drogas, você pode esconder a droga *Y*. No menu **Visualizar** , escolha **Modo de Editar**, em seguida, clique com o botão direito do mouse sobre o ponto *Y* da droga e escolha **Hide e Replan**.

Na trama simplificada, o medicamento *Y* e todos os seus links estão escondidos. Agora, é possível ver claramente que apenas as drogas *A* e *B* estão associadas à pressão alta. Apenas as drogas *C* e *X* estão associadas a baixa pressão arterial. E a pressão arterial normal está associada apenas com o medicamento *X*. A esta altura, porém, você ainda não sabe como escolher entre os medicamentos *A* e *B* ou entre os medicamentos *C* e *X*, para um determinado paciente. É aqui que a modelagem pode ajudar.

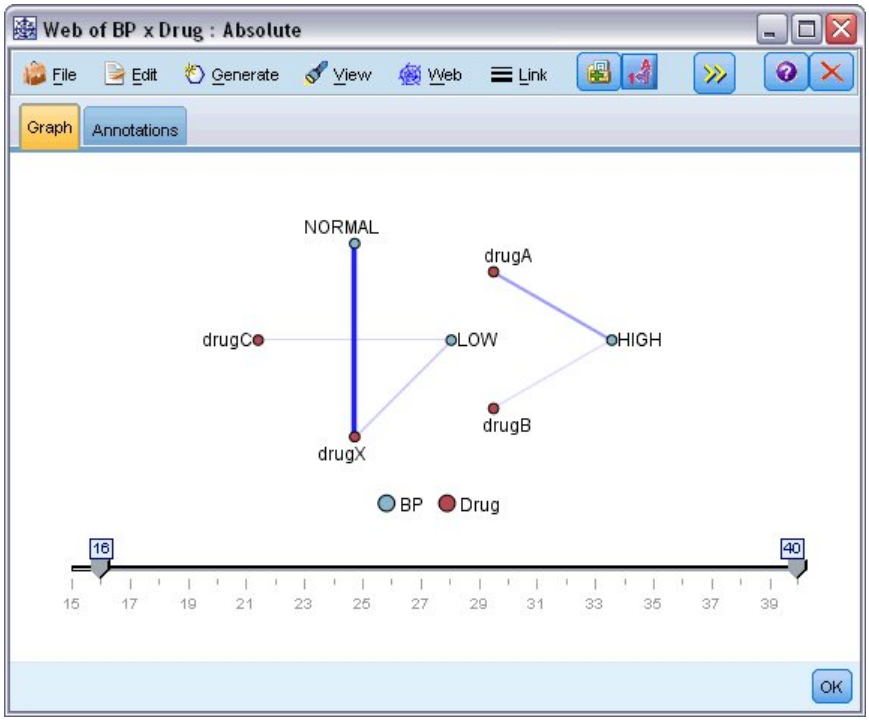


Figura 89. Gráfico web com droga Y e seus links escondidos

Derivando um novo campo

Uma vez que a proporção de sódio para potássio parece predizer quando usar o medicamento *Y*, é possível derivar um campo que contém o valor dessa proporção para cada registro. Este campo pode ser útil posteriormente, quando você construir um modelo para prever quando usar cada um dos cinco medicamentos. Para simplificar o layout do fluxo, inicie excluindo todos os nós, exceto o nó de origem DRUG1n . Conecte um nó de Derivação (guia Ops de campo) a DRUG1n, em seguida, dê um duplo clique no nó Derivar para editá-lo.

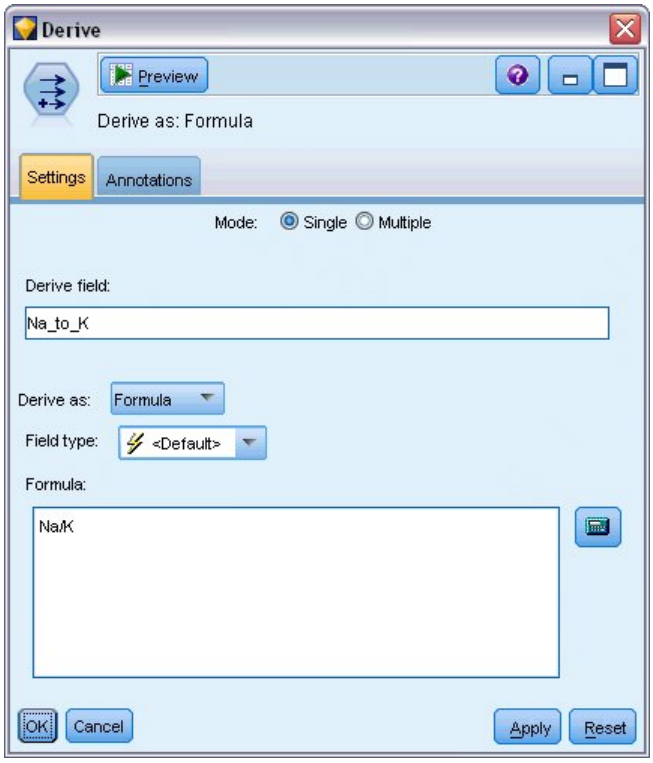


Figura 90. Editando o Nó Derivar

Nomeie o novo campo *Na_to_K*. Como você obtém o novo campo dividindo o valor de sódio pelo valor de potássio, insira Na/K para a fórmula. Você também pode criar uma fórmula clicando no ícone apenas para a direita do campo. Isso abre o construtor de expressões, uma maneira de criar expressões interativamente usando listas internas de funções, operandos e campos e seus valores.

É possível verificar a distribuição de seu novo campo anexando um nó Histograma ao nó Derivar. Na caixa de diálogo do nó do Histograma, especifique *Na_to_K* como o campo a ser plotado e *Drug* como o campo de sobreposição.

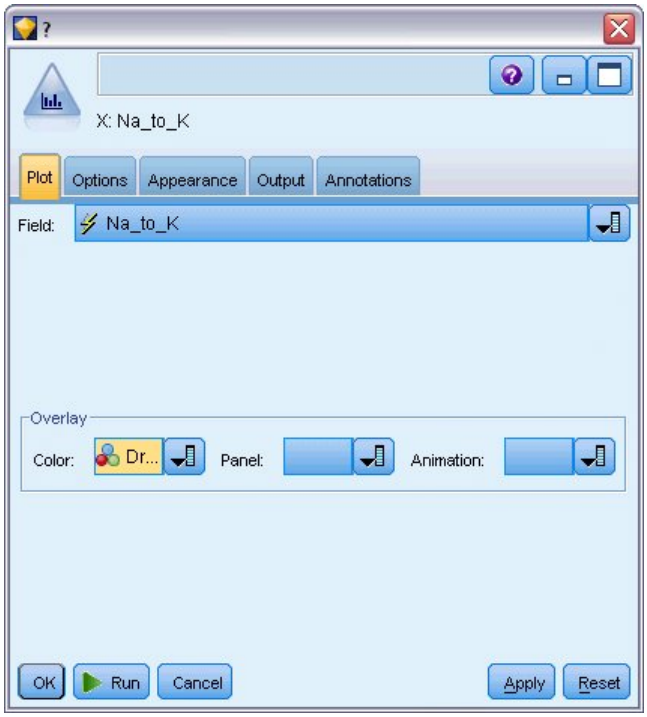


Figura 91. Editando o nó do Histograma

Quando você executa o fluxo, você recebe o gráfico mostrado aqui. Com base no display, pode-se concluir que quando o valor *Na_to_K* for cerca de 15 ou acima, o medicamento *Y* é a droga de escolha.

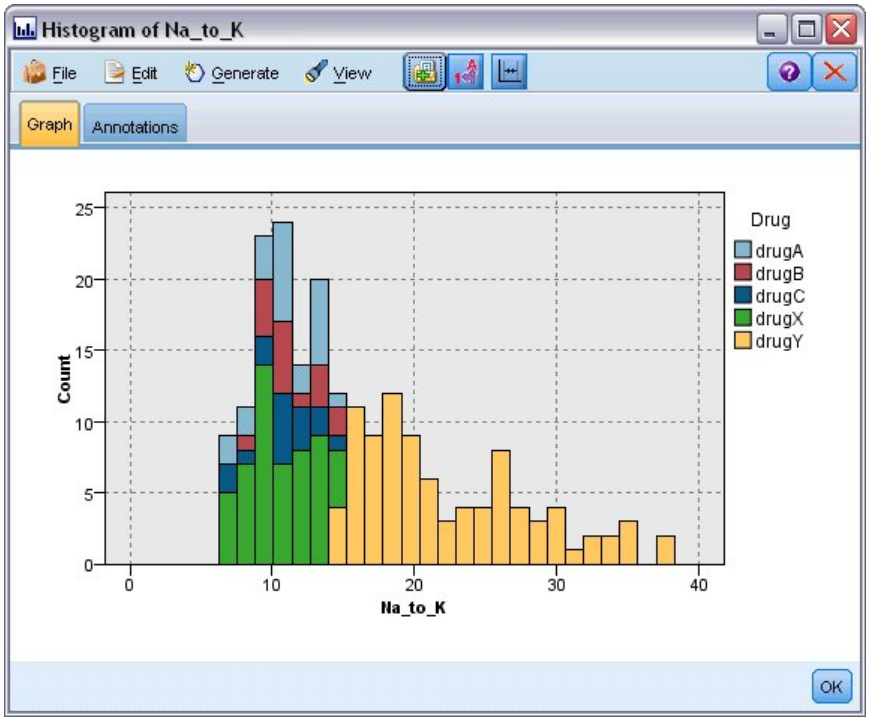


Figura 92. Exibição de histograma

Construindo um modelo

Ao explorar e manipular os dados, você foi capaz de formar algumas hipóteses. A proporção de sódio para potássio no sangue parece afetar a escolha do medicamento, assim como a pressão arterial. Mas não é possível explicar totalmente todos os relacionamentos ainda. É aqui que a modelagem provavelmente fornecerá algumas respostas. Neste caso, você usará try para ajustar os dados usando um modelo de construção de regras, C5.0.

Já que você está usando um campo derivado, *Na_to_K*, você pode filtrar os campos originais, *Na* e *K*, para que eles não sejam usados duas vezes no algoritmo de modelagem. Você pode fazer isso usando um nó Filtro.

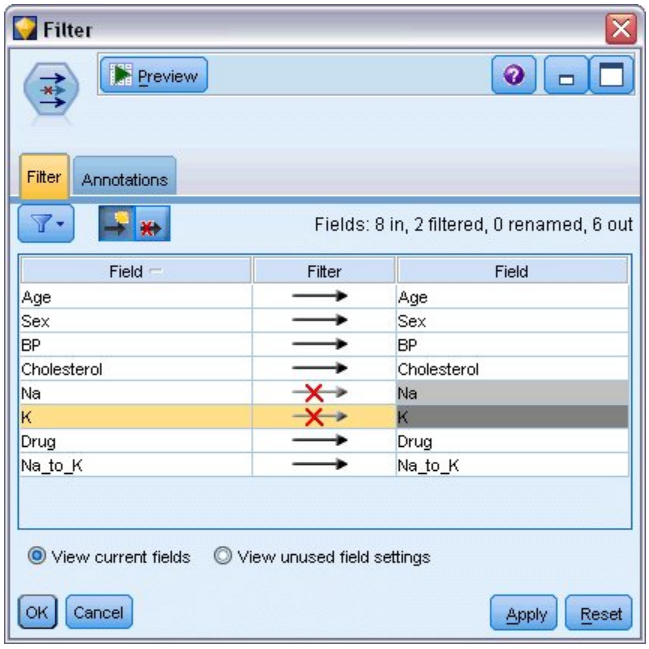


Figura 93. Editando o Nó Filtro

Na guia Filtro, clique nas setas ao lado de *Na* e *K*. Os Xs vermelhos aparecem sobre as setas para indicar que os campos agora estão filtrados.

Em seguida, anexe um nó Tipo conectado ao nó Filtro. O nó Type permite que você indique os tipos de campos que você está usando e como eles são usados para prever os resultados.

Na guia Tipos, configure a função para o campo *Drug* para **Target**, indicando que *Drug* é o campo que você deseja prever. Deixe a função para os outros campos configurados para **Entrada** para que eles sejam usados como preditores.

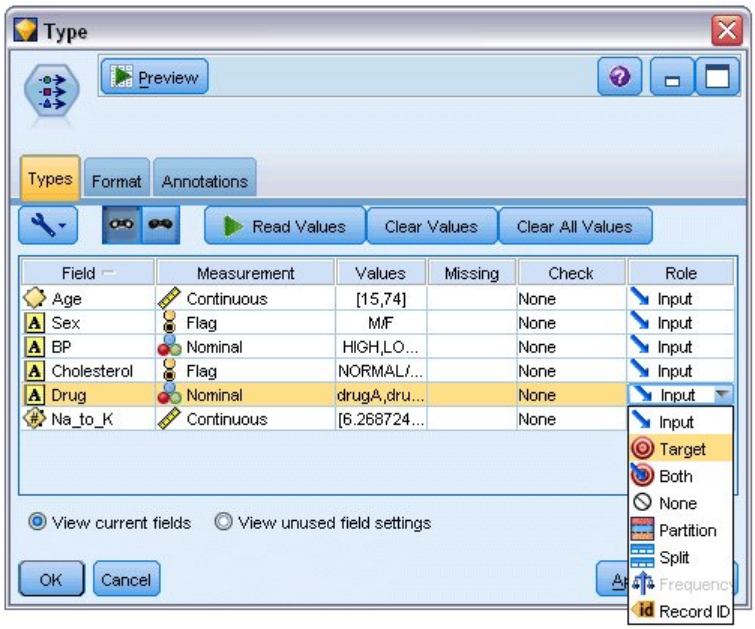


Figura 94. Editando o nó Tipo

Para estimar o modelo, coloque um nó C5.0 na área de trabalho e anexe-a até o final do fluxo conforme mostrado. Em seguida, clique no botão da barra de ferramentas verde **Executar** para executar o fluxo.

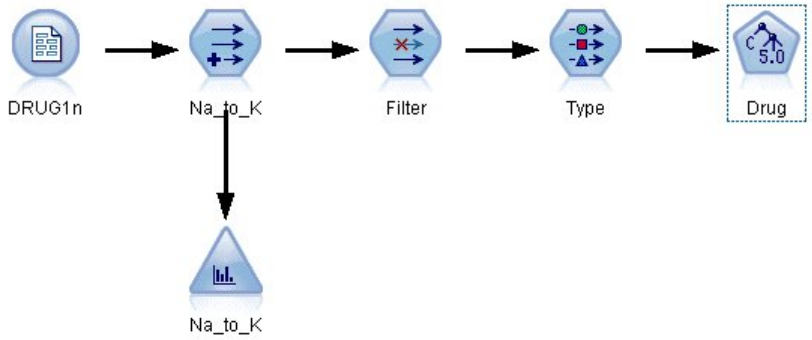


Figura 95. Adicionando um nó C5.0

Navegando no modelo

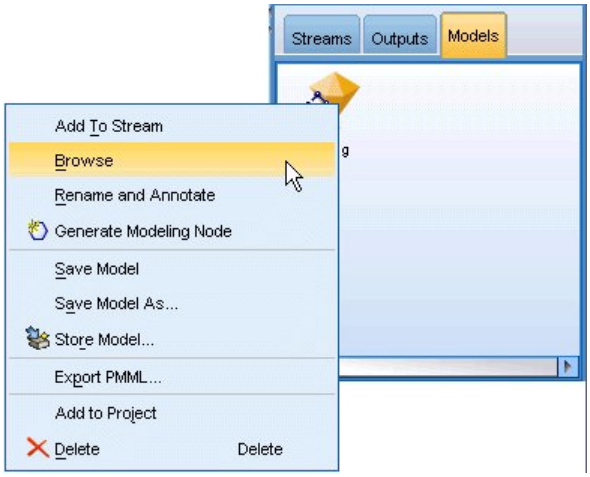


Figura 96. Navegando no modelo

Quando o nó C5.0 é executado, o nugget modelo é adicionado ao fluxo e também à paleta de Models no canto superior direito da janela. Para navegar no modelo, clique com o botão direito do mouse em qualquer um dos ícones e escolha **Editar** ou **Procurar** no menu de contexto.

O navegador de Regra exibe o conjunto de regras geradas pelo nó C5.0 em um formato de árvore de decisão. Inicialmente, a árvore está desmoronada. Para expanda-lo, clique no botão **Todos** para mostrar todos os níveis.

Figura 97. Navegador de regra

Agora é possível ver as peças que faltam no quebra-cabeça. Para pessoas com uma proporção de *Na*-para-*K* menor que 14.64 e pressão arterial alta, a idade determina a escolha da droga em questão. Para pessoas com pressão arterial baixa, o nível de colesterol parece ser o melhor preditor.

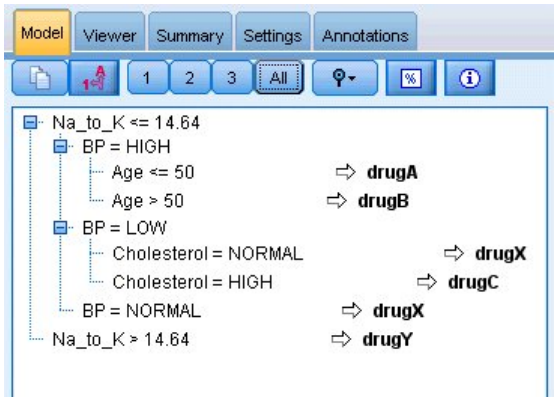


Figura 98. Navegador de regra totalmente expandido

A mesma árvore de decisão pode ser visualizada em um formato gráfico mais sofisticado, clicando na aba **Viewer** . Aqui, é possível ver com mais facilidade o número de casos para cada categoria de pressão arterial, assim como o percentual de casos.

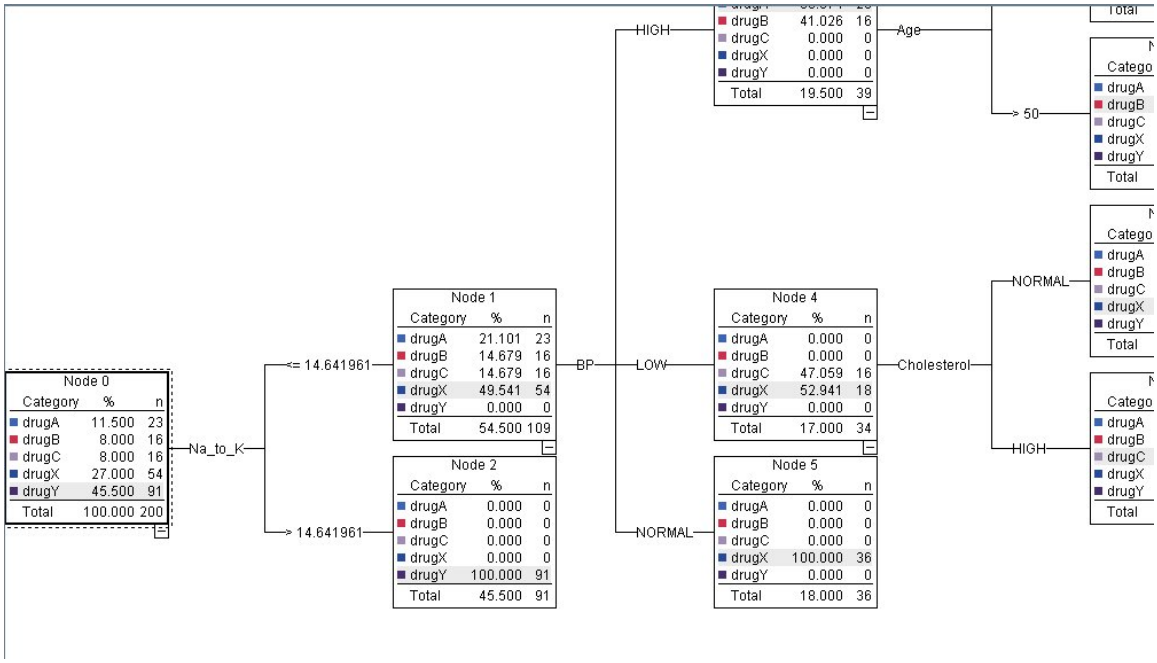


Figura 99. Árvore de decisão em formato gráfico

Usando um nó de análise

É possível avaliar a precisão do modelo usando um nó de análise. Conecte um nó de Análise (da paleta do nó de saída) ao nugget do modelo, abra o nó Análise e clique em **Executar**.

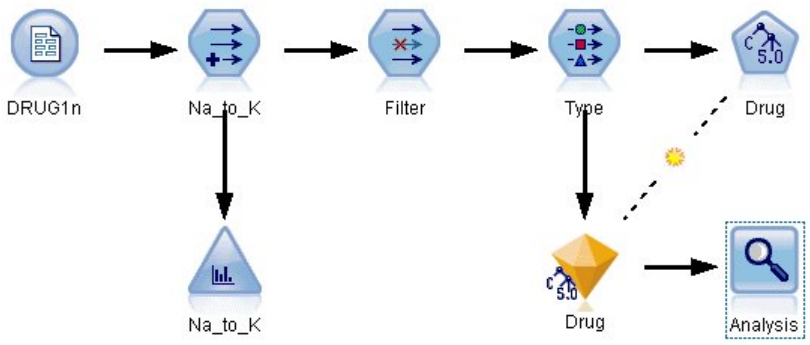


Figura 100. Adicionando um Nó de Análise

A saída do nó de análise mostra que, com esse conjunto de dados artificial, o modelo previu corretamente a escolha do medicamento para cada registro no conjunto de dados. Com um conjunto de dados real, é improvável que você veja 100% de precisão, mas pode usar o nó de análise para ajudar a determinar se o modelo é aceitavelmente preciso para seu aplicativo específico.

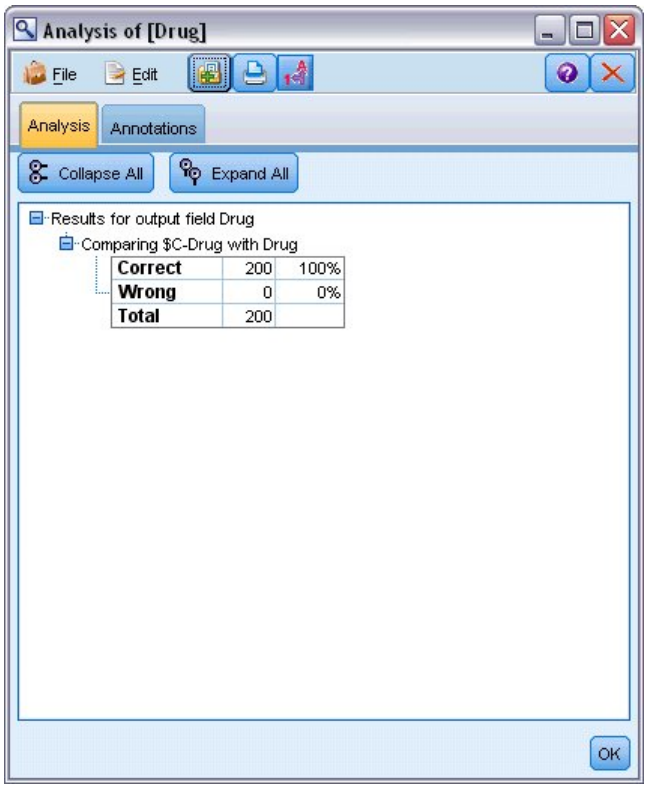


Figura 101. Saída do nó de análise

Capítulo 9. Preditores De Triagem (Seleção de Recurso)

O nó da Seleção de Recursos ajuda você a identificar os campos que são mais importantes na previsão de um determinado resultado. A partir de um conjunto de centenas ou mesmo milhares de preditores, o nó Seleção de Recursos faz a triagem, classifica e seleciona os preditores que podem ser mais importantes. Em última análise, você pode acabar com um modelo mais rápido, mais eficiente-um que usa menos preditores, executa mais rapidamente, e pode ser mais fácil de entender.

Os dados utilizados neste exemplo representam um armazém de dados para uma empresa de telefonia hipotética e contêm informações sobre respostas a uma promoção especial por parte de 5.000 dos clientes da empresa. Os dados incluem um grande número de campos contendo as estatísticas de idade, emprego, renda e uso de telefone dos clientes. Três campos de "destino" mostram se o cliente respondeu ou não a cada uma das três ofertas. A empresa deseja usar esses dados para ajudar a prever quais clientes têm maior probabilidade de responder a ofertas semelhantes no futuro.

Este exemplo usa o fluxo denominado *featureselection.str*, que faz referência ao arquivo de dados denominado *customer_dbase.sav* Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *featureselection.str* está no diretório *streams* .

Este exemplo se concentra em apenas uma das ofertas como destino. Ele usa o nó de construção de árvore CHAID para desenvolver um modelo para descrever quais clientes têm maior probabilidade de responder à promoção. Ele contrasta duas abordagens:

- Sem seleção de recurso. Todos os campos preditores no conjunto de dados são usados como entradas para a árvore CHAID.
- Com seleção de recurso. O nó Seleção de Recurso é usado para selecionar os 10 principais preditores. Eles são então inseridos na árvore CHAID.

Ao comparar os dois modelos de árvore resultantes, podemos ver como a seleção de recursos produz resultados efetivos.

Construindo o Fluxo

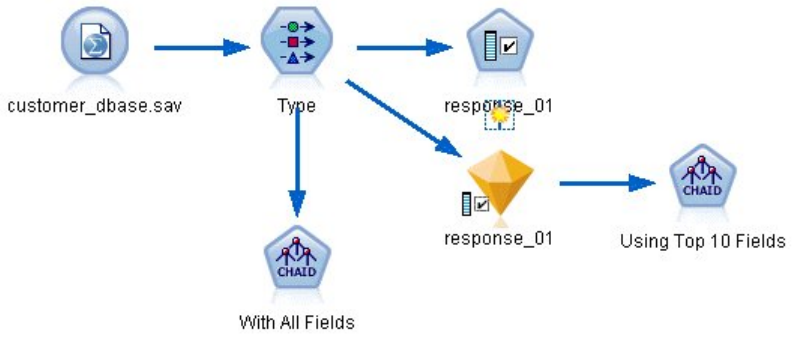


Figura 102. Fluxo de exemplo de recurso Seleção

1. Coloque um nó de origem do Arquivo de Estatísticas sobre uma tela de fluxo em branco. Aponte este nó para o arquivo de dados de exemplo *customer_dbase.sav*, disponível no diretório *Demos* em sua instalação do IBM SPSS Modelador (Como alternativa, abra o arquivo de fluxo de exemplo *featureselection.str* no diretório *streams* .).
2. Adicionar um nó Tipo. Na guia Tipos, role até a parte inferior e altere a função para *response_01* para *Target*. Altere a função para *Nenhum* para os outros campos de resposta (*response_02* e

response_03) assim como para o ID do cliente (*custid*) no topo da lista. Deixe o papel configurado para *Entrada* para todos os outros campos, e clique no botão **Ler Valores** , em seguida, clique em **OK**.

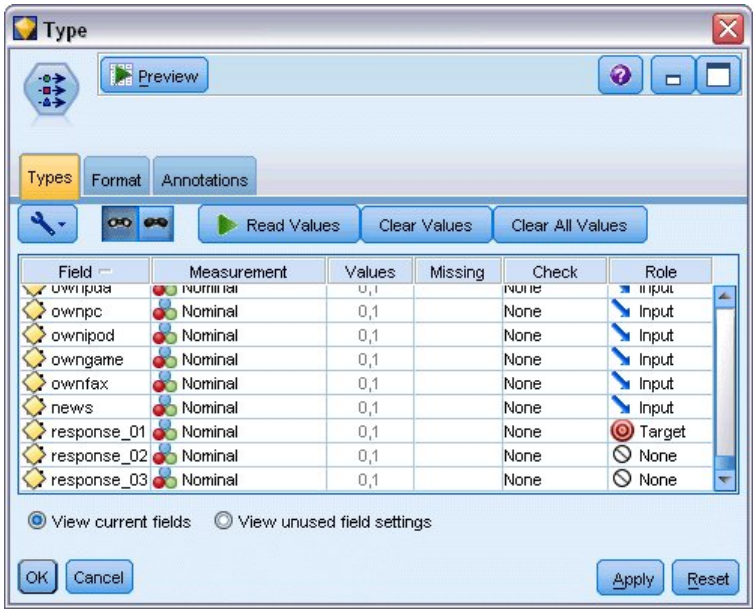


Figura 103. Incluindo um Nó de Tipo

3. Inclua um nó de modelagem de Seleção de Recurso no fluxo. Neste nó, é possível especificar as regras e critérios para triagem, ou desqualificar, campos.
4. Execute o fluxo para criar a nugget do modelo de Seleção de Recursos.
5. Clique com o botão direito do mouse no nugget modelo no fluxo ou na paleta de Modelos e escolha **Editar** ou **Procurar** para observar os resultados.

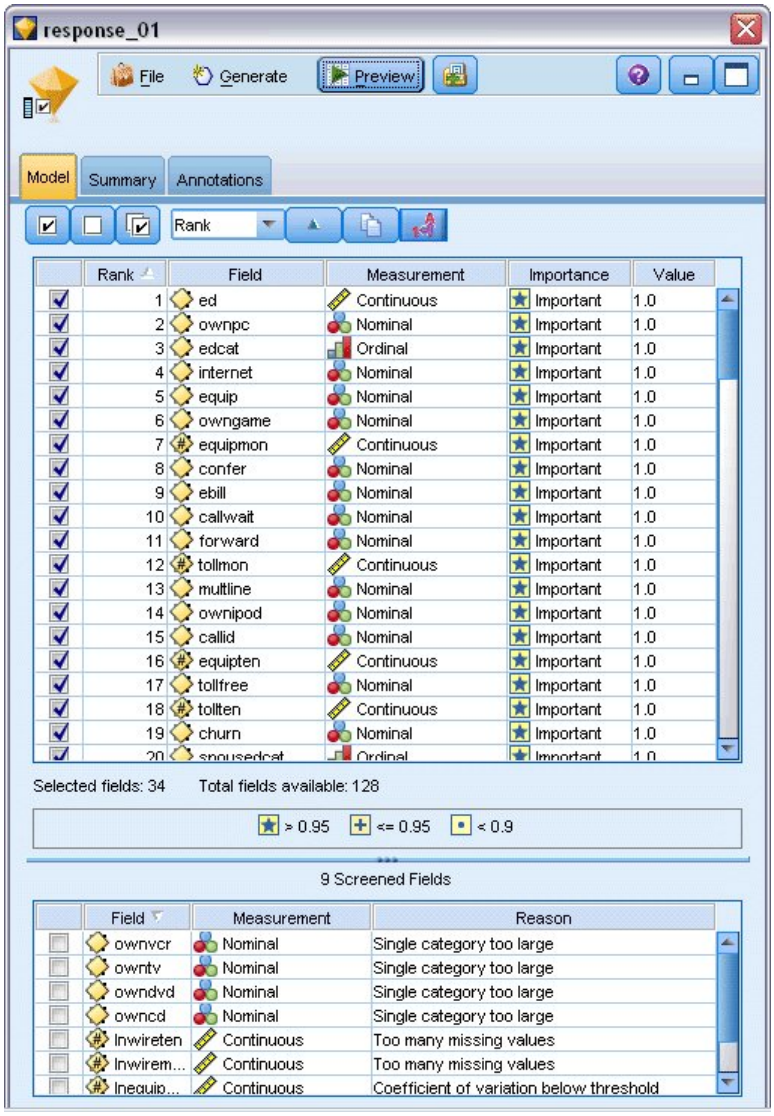


Figura 104. Guia do modelo em modelo de Seleção de Recursos nugget

O painel superior mostra os campos encontrados para serem úteis na predição. Estes são ranqueados com base em importância. O painel inferior mostra quais campos foram gritados a partir da análise e por quê. Ao examinar os campos no painel superior, é possível decidir quais são os que utilizar em sessões de modelagem subsequentes.

6. Agora podemos selecionar os campos para usar downstream. Apesar de 34 campos terem sido originalmente identificados como importantes, queremos reduzir ainda mais o conjunto de preditores.
7. Selecione apenas os 10 principais preditores usando as marcas de verificação na primeira coluna para deseleger os preditores indesejados. (Clique na marca de verificação na linha 11, segure a tecla Shift e clique na marca de verificação na linha 34.) Feche o nugget modelo.
8. Para comparar resultados sem seleção de recursos, você deve adicionar dois nós de modelagem CHAID ao fluxo: um que usa seleção de recursos e outro que não.
9. Conecte um nó CHAID ao nó do Tipo, e o outro ao nugget modelo de Seleção de Recurso.
10. Abra cada nó CHAID, selecione a guia Opções de Construção e certise-se de que as opções **Construir novo modelo**, **Construir uma única árvore** e **Ativar sessão interativa** são selecionadas na pane de Objetivos.

Na pane Basics, certise-se de que **Máximo de Profundidade de Árvore** esteja configurado para 5.

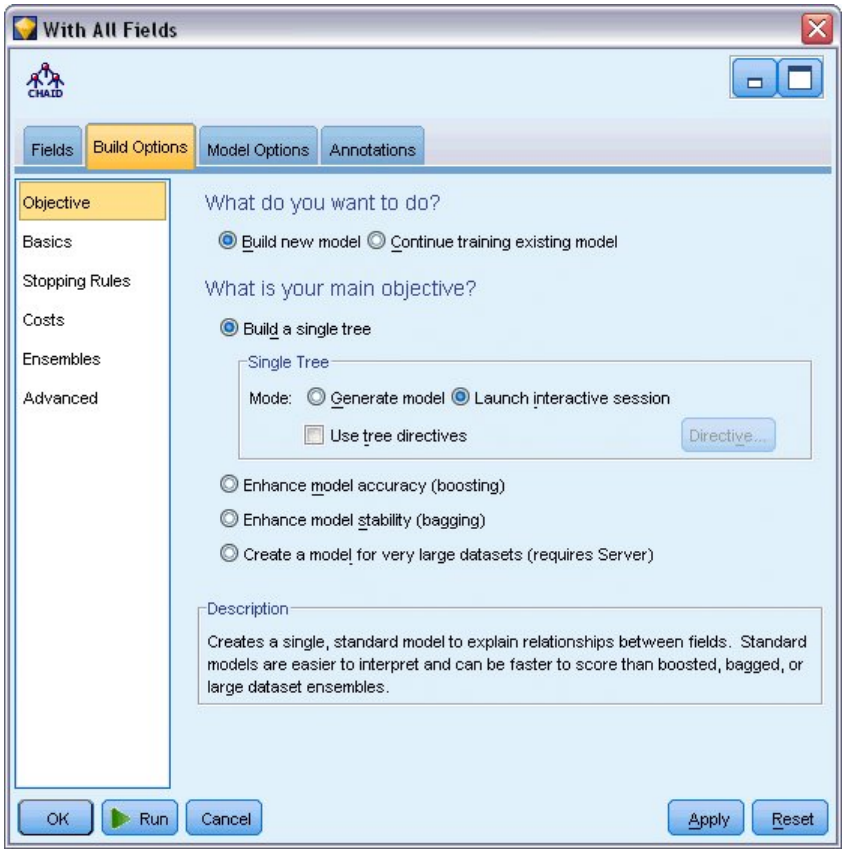


Figura 105. Configurações de objetivos para o nó de modelagem CHAID para todos os campos do preditor

Construindo os modelos

1. Execute o nó CHAID que usa todos os preditores no dataset (aquele conectado ao nó Type). À medida que corre, observe quanto tempo leva para executar. A janela de resultados exibe uma tabela.
2. A partir dos menus, escolha **Tree> Grow Tree** para crescer e exibir a árvore expandida.

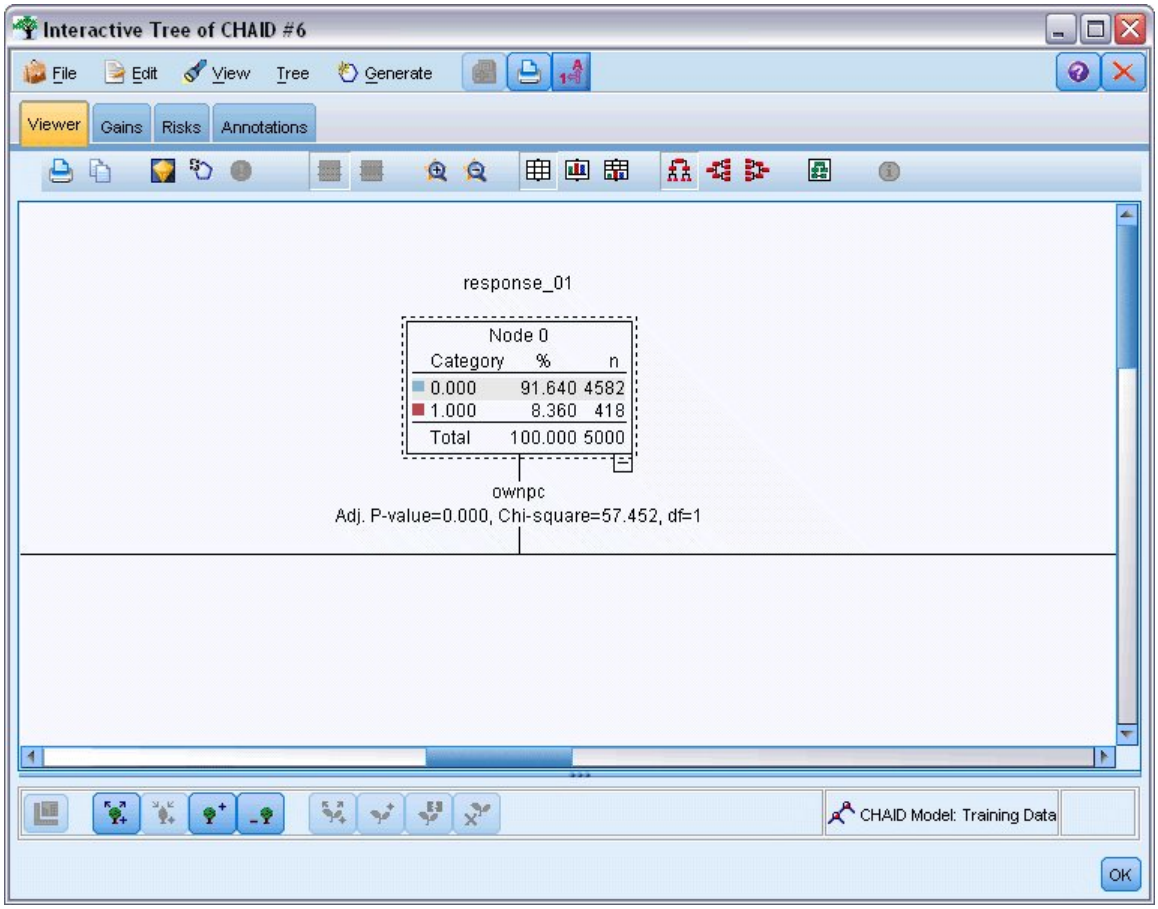


Figura 106. Crescendo a árvore no Tree Builder

3. Agora faça o mesmo para o outro nó CHAID, que usa apenas 10 preditores. Novamente, crese a árvore quando o Tree Builder abrir.

O segundo modelo deveria ter executado mais rápido do que o primeiro. Como esse dataset é razoavelmente pequeno, a diferença nos tempos de execução é provavelmente de alguns segundos; mas para maiores datasets do mundo real, a diferença pode ser muito perceptível-minutos ou até mesmo horas. Usar a seleção de recursos pode acelerar drasticamente o tempo de processamento.

A segunda árvore também contém menos nós da árvore do que a primeira. É mais fácil compreender. Mas antes de decidir utilizá-lo, é preciso descobrir se ele é eficaz e como ele se compara ao modelo que usa todos os preditores.

Comparando os Resultados

Para comparar os dois resultados, precisamos de uma medida de eficácia. Para isso, utilizaremos a guia Gains no Tree Builder. Vamos olhar para **lift**, que mede o quanto mais prováveis os registros em um nó devem cair sob a categoria de destino quando comparados a todos os registros no dataset. Por exemplo, um valor de elevação de 148% indica que registros no nó são 1.48 vezes mais propensos a cair na categoria de destino do que todos os registros no conjunto de dados. Levantamento é indicado na coluna *Índice* na guia Gains.

1. No Tree Builder para o conjunto completo de preditores, clique na guia Gains. Altere a categoria de destino para 1.0 Altere a exibição para quartiles clicando primeiro no botão da barra de ferramentas Quantiles. Em seguida, selecione **Quartile** na lista suspensa à direita deste botão.
2. Repita esse procedimento no Tree Builder para o conjunto de 10 preditores para que você tenha duas tabelas de Gains semelhantes para comparar, como mostrado nas figuras a seguir.

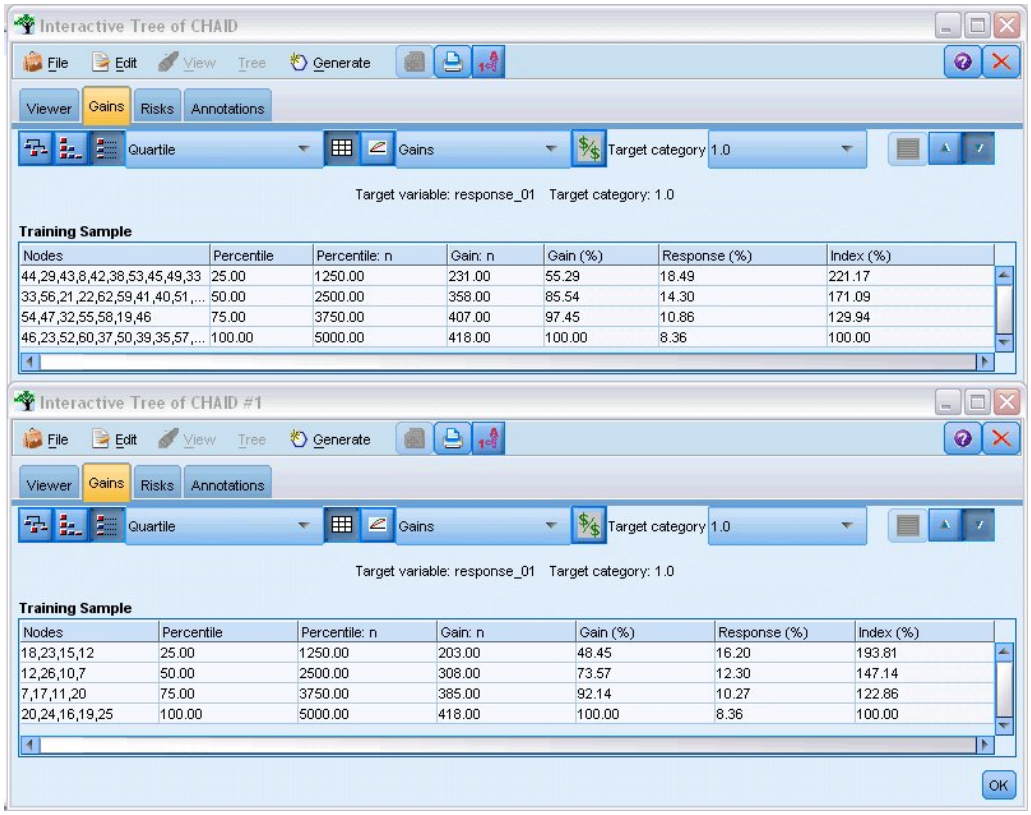


Figura 107. Gráficos de ganhos para os dois modelos CHAID

Cada Tabela Ganhos agrupa os nós terminais para sua árvore em quartiles. Para comparar a eficácia dos dois modelos, veja o levantamento (valor de*Índice*) para o quartil superior em cada tabela.

Quando todos os preditores estão incluídos, o modelo mostra um elevador de 221%. Ou seja, os casos com as características nesses nós são 2.2 vezes mais propensos a responder à promoção de destino Para ver quais são essas características, clique para selecionar a linha superior. Em seguida, alterna para a guia Viewer, onde os nós correspondentes são agora delineados em preto. Siga a árvore abaixo até cada nó do terminal destacado para ver como os preditores foram divididos. O quartil superior sozinho inclui 10 nós. Quando traduzidos em modelos de pontuação do mundo real, 10 perfis de clientes diferentes podem ser difíceis de gerenciar.

Com apenas os 10 principais preditores (conforme identificados por seleção de recursos) incluídos, o elevador é de quase 194%. Embora esse modelo não seja tão bom quanto o modelo que usa todos os preditores, ele certamente é útil. Aqui, o quartil superior inclui apenas quatro nós, portanto, é mais simples. Por isso, podemos determinar que o modelo de seleção de recursos seja preferível a aquele com todos os preditores.

Resumo

Vamos rever as vantagens da seleção de recursos. Usar menos preditores é menos caro. Significa que você tem menos dados para coletar, processar e alimentar em seus modelos. O tempo de cálculo foi aprimorado. Neste exemplo, mesmo com a etapa de seleção de recurso extra, o prédio do modelo foi perceptivelmente mais rápido com o conjunto menor de preditores. Com um conjunto de dados maior do mundo real, a economia de tempo deve ser bastante ampliada.

Usar menos preditores resulta em uma pontuação mais simples. Como mostra o exemplo, você pode identificar apenas quatro perfis de clientes que provavelmente respondirão à promoção. Observe que, com um número maior de preditores, você corre o risco de super ajustar seu modelo. O modelo mais simples pode generalizar melhor para outros conjuntos de dados (embora você precise testar isso para ter certeza).

Você poderia ter usado um algoritmo de construção de árvore para fazer o trabalho de seleção de recursos, permitindo que a árvore identifique os preditores mais importantes para você. Na verdade, o algoritmo CHAID é frequentemente usado para esta finalidade, e é mesmo possível fazer crescer o nível de árvore-por-nível para controlar sua profundidade e complexidade. No entanto, o nó de Seleção de Recursos é mais rápido e mais fácil de usar. Ele classifica todos os preditores em uma etapa rápida, permitindo que você identifique os campos mais importantes rapidamente. Ele também permite que você varie o número de preditores a incluir. Você poderia facilmente executar este exemplo novamente usando os 15 ou 20 preditores superiores em vez de 10, comparando os resultados para determinar o modelo ideal.

Capítulo 10. Reduzindo Comprimento De Cadeia De Dados De Entrada (Nó Reclassificfy)

Reduzindo Comprimento De Cadeia De Dados De Entrada (Reclassifique)

Para regressão logística binomial e modelos de classificador automático que incluem um modelo de regressão logística binomial, os campos de sequência de caracteres são limitados a um máximo de oito caracteres. Onde strings são mais de oito caracteres, eles podem ser recodificados usando um nó Reclassify.

Este exemplo usa o fluxo denominado *reclassify_strings.str*, que faz referência ao arquivo de dados denominado *drogar_long_name*.. Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *reclassify_strings.str* está no diretório *streams*

Este exemplo foca em uma pequena parte de um fluxo para mostrar o tipo de erros que podem ser gerados com strings overlong e explica como utilizar o nó Reclassify para alterar os detalhes da string para um comprimento aceitável. Embora o exemplo use um nó de regressão logística binomial, é igualmente aplicável ao usar o nó de classificação automática para gerar um modelo de regressão logística binomial.

Reclassificando os dados

1. Usando um nó de origem do Arquivo Variável, conecte-se ao dataset *drogar_long_name* na pasta *Demos* .

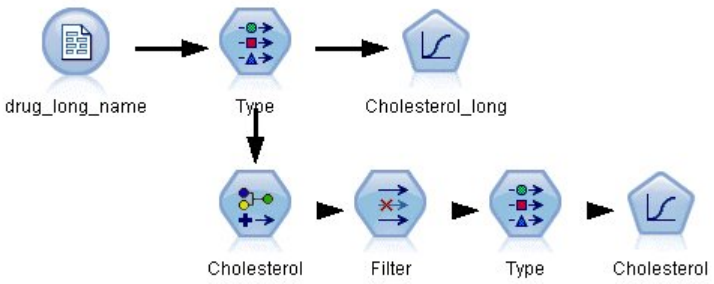


Figura 108. Amostra de fluxo mostrando reclassificação de cadeia para regressão logística binomial

2. Inclua um nó Tipo no nó Fonte e selecione **Cholesterol_long** como o destino.
3. Inclua um nó de Regressão Logística no nó Type.
4. No nó da Regressão Logística, clique na guia Modelo e selecione o procedimento **Binomial** .

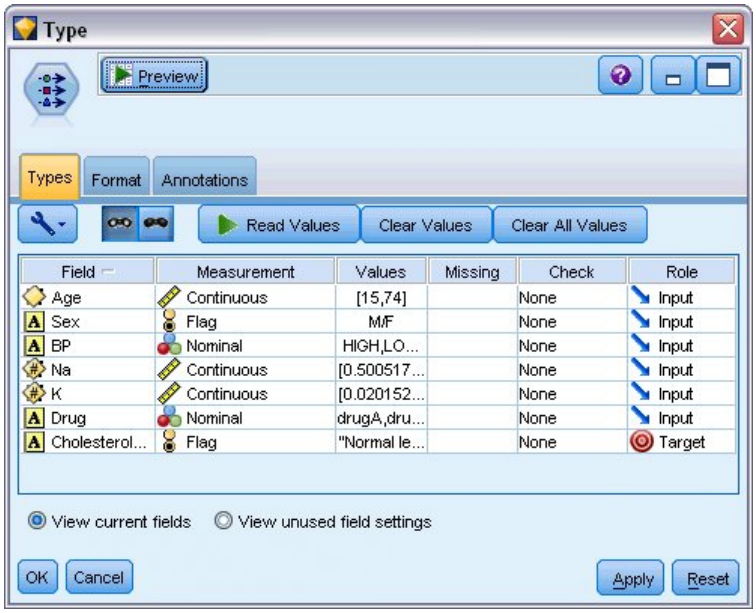


Figura 109. Detalhes de cadeia longa no campo "Cholesterol_long"

5. Ao executar o nó de Regressão Logística em *reclassify_strings.str*, uma mensagem de erro é exibida avisando que os valores de sequência **Cholesterol_long** são muito longos.

   Se você encontrar este tipo de mensagem de erro, siga o procedimento explicado no restante deste exemplo para modificar seus dados.

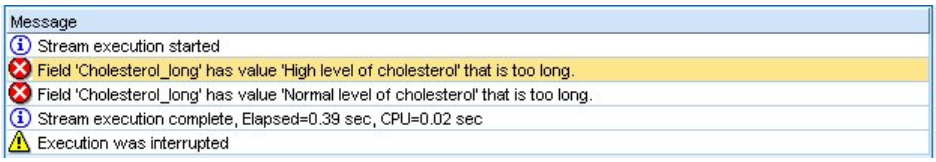


Figura 110. Mensagem de erro exibida ao executar o nó de regressão logístico binomial

6. Adicio um nó Reclassify ao nó Type.
7. No campo Reclassify, selecione **Cholesterol_long**.
8. Digite **Cholesterol** como o novo nome de campo.
9. Clique no botão **Obter** para adicionar os valores **Cholesterol_long** à coluna de valor original.
10. Na nova coluna de valor, digite **High** próximo ao valor original de **High level of cholesterol** e **Normal** próximo ao valor original de **Normal level of cholesterol**.

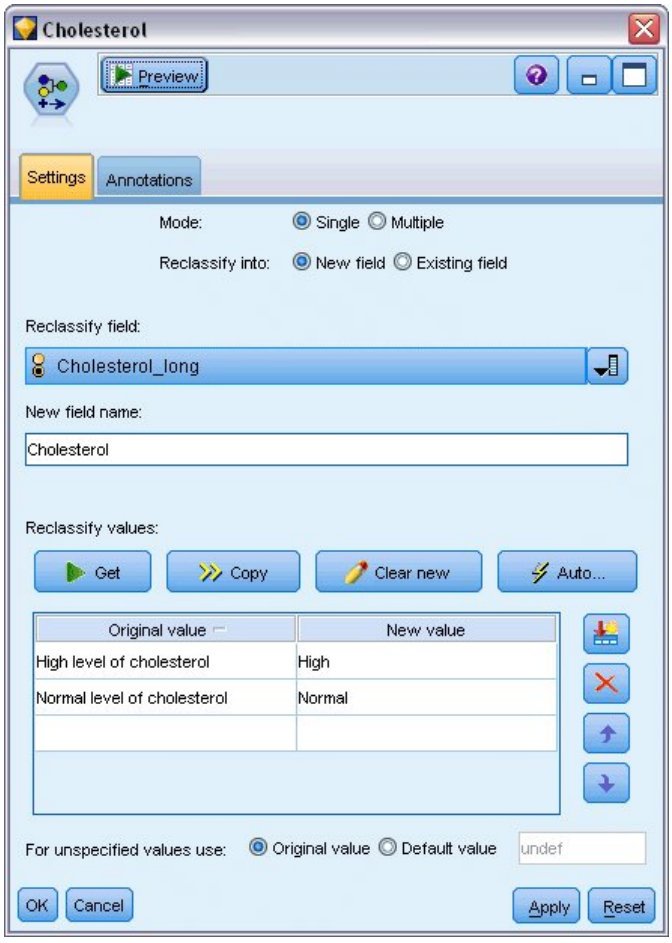


Figura 111. Reclassificando as cordas longas

11. Inclua um nó Filtro no nó Reclassify.
12. Na coluna Filtro, clique para remover **Cholesterol_long**.

Figura 112. Filtrando o campo "Cholesterol_long" dos dados

13. Adiciona um nó Tipo ao nó Filtro e selecione **Cholesterol** como o destino.

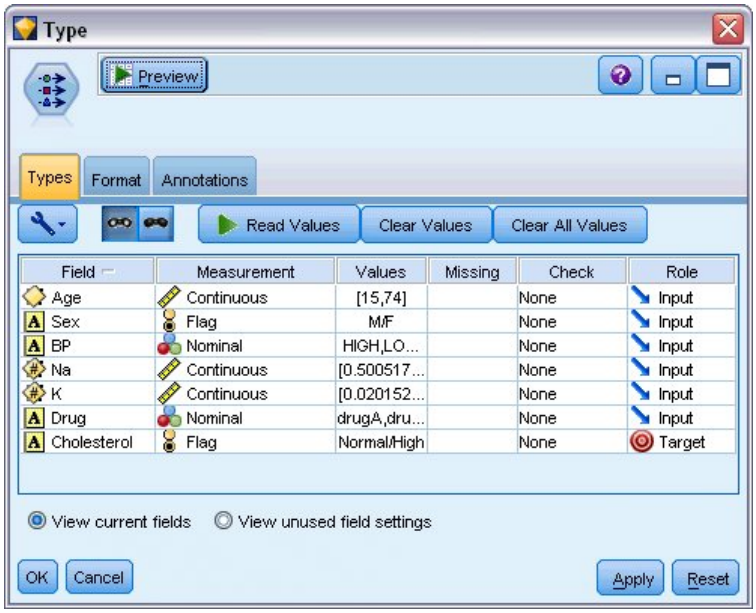


Figura 113. Detalhes da sequência de caracteres curta no campo "Cholesterol"

14. Inclua um Nó Logístico no nó Type.
15. No nó Logístico, clique na guia Modelo e selecione o procedimento **Binomial** .
16. Agora é possível executar o nó Logístico Binomial e gerar um modelo sem exibir uma mensagem de erro.

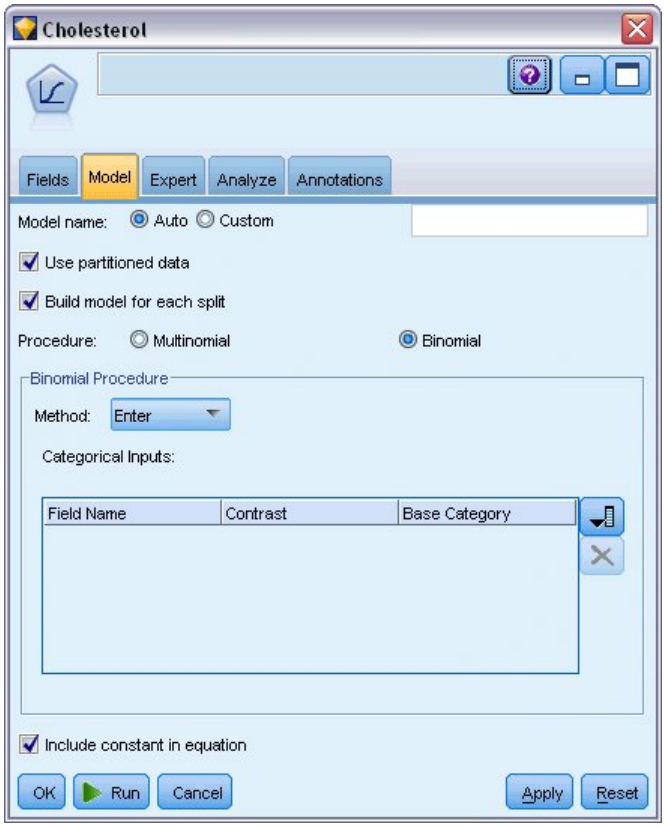


Figura 114. Escolhendo Binomial como procedimento

Este exemplo só mostra parte de um fluxo. Se você precisar de mais informações sobre os tipos de fluxos nos quais você pode precisar reclassificar as cadeias longas, os exemplos a seguir estão disponíveis:

- Nó do Classificador Automático. Consulte o tópico “Modelagem de resposta do cliente (classificador automático)” na página 33 para obter informações adicionais.
- Nó Binomial Logistic Regression. Consulte o tópico Capítulo 13, “Churn de Telecomunicações (Regressão Logística Binomial)”, na página 129 para obter informações adicionais.

Mais informações sobre como utilizar IBM SPSS Modelador, como guia de um usuário, referência do nó e guia de algoritmos, estão disponíveis a partir do diretório *Documentação* do disco de instalação.

Capítulo 11. Modelagem de Resposta ao Cliente (Lista de Decisão)

O algoritmo Lista de Decisão gera regras que indicam uma probabilidade maior ou menor de um dado resultado binário (sim ou não). Os modelos de Lista de Decisão são amplamente utilizados na gestão de relacionamento com o cliente, como aplicativos de call center ou de marketing.

Este exemplo é baseado em uma empresa fictícia que deseja obter resultados mais rentáveis em futuras campanhas de marketing, combinando a oferta correta a cada cliente. Especificamente, o exemplo usa um modelo de Lista de Decisão para identificar as características dos clientes que são mais propensos a responder favoravelmente, com base em promoções anteriores, e a gerar uma lista de mailing com base nos resultados.

Os modelos de Lista de Decisão são particularmente bem adequados à modelagem interativa, permitindo ajustar parâmetros no modelo e ver imediatamente os resultados. Para uma abordagem diferente que permite criar automaticamente uma série de modelos diferentes e classificar os resultados, o nó Auto Classifier pode ser usado em vez disso.

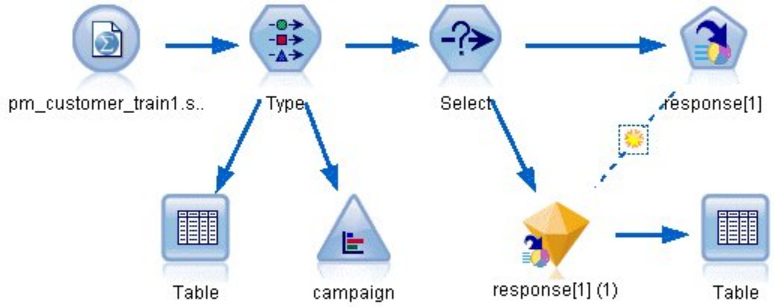
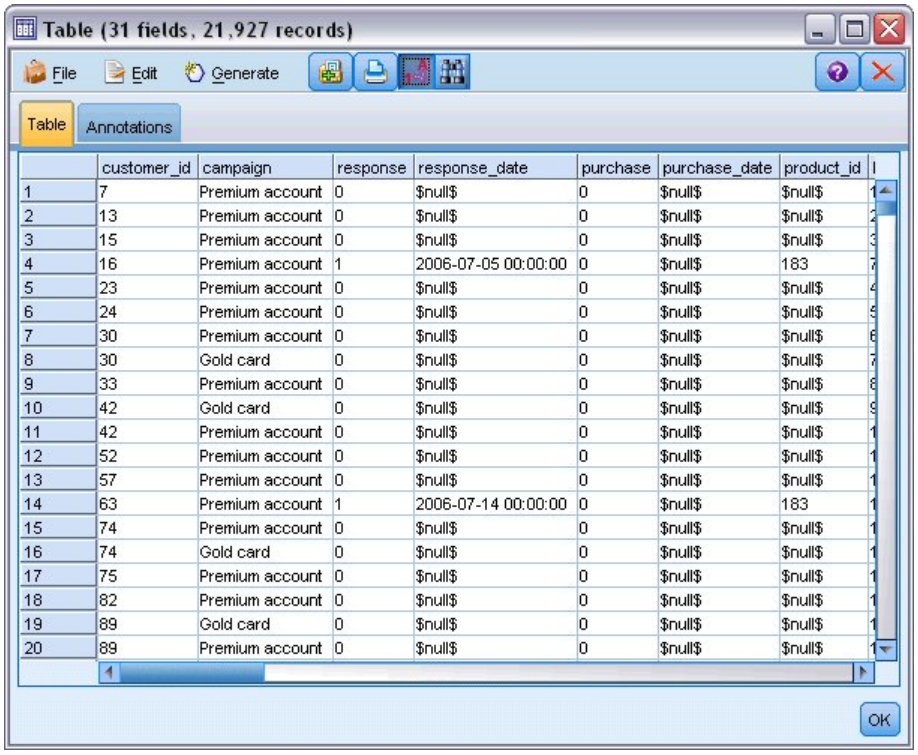


Figura 115. Fluxo de amostra Lista de decisão

Este exemplo usa o fluxo *pm_decisionlist.str*, que faz referência ao arquivo de dados *pm_customer_train1.sav* Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *pm_decisionlist.str* está no diretório *stream* .

Dados históricos

O arquivo *pm_customer_train1.sav* tem dados históricos rastreando as ofertas feitas a clientes específicos em campanhas passadas, conforme indicado pelo valor do campo *campanha* . O maior número de registros cai na campanha de *conta premium*.



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id
1	7	Premium account	0	$null$	0	$null$	$null$
2	13	Premium account	0	$null$	0	$null$	$null$
3	15	Premium account	0	$null$	0	$null$	$null$
4	16	Premium account	1	2006-07-05 00:00:00	0	$null$	183
5	23	Premium account	0	$null$	0	$null$	$null$
6	24	Premium account	0	$null$	0	$null$	$null$
7	30	Premium account	0	$null$	0	$null$	$null$
8	30	Gold card	0	$null$	0	$null$	$null$
9	33	Premium account	0	$null$	0	$null$	$null$
10	42	Gold card	0	$null$	0	$null$	$null$
11	42	Premium account	0	$null$	0	$null$	$null$
12	52	Premium account	0	$null$	0	$null$	$null$
13	57	Premium account	0	$null$	0	$null$	$null$
14	63	Premium account	1	2006-07-14 00:00:00	0	$null$	183
15	74	Premium account	0	$null$	0	$null$	$null$
16	74	Gold card	0	$null$	0	$null$	$null$
17	75	Premium account	0	$null$	0	$null$	$null$
18	82	Premium account	0	$null$	0	$null$	$null$
19	89	Gold card	0	$null$	0	$null$	$null$
20	89	Premium account	0	$null$	0	$null$	$null$

Figura 116. Dados sobre promoções anteriores

Os valores do campo *campanha* são realmente codificados como números inteiros nos dados, com etiquetas definidas no nó Type (por exemplo, *2 = Conta Premium*). Você pode alternar exibição de etiquetas de valor na tabela usando a barra de ferramentas.

O arquivo também inclui uma série de campos contendo informações demográficas e financeiras sobre cada cliente que podem ser usados para construir ou "treinar" um modelo que prevê taxas de resposta para diferentes grupos com base em características específicas.

Construindo o Fluxo

1. Adicione um nó do Arquivo de Estatísticas apontando para *pm_customer_train1.sav*, localizado na pasta *Demos* de sua instalação IBM SPSS Modelador . (Você pode especificar `$CLEO_DEMOS/` no caminho de arquivo como um atalho para referencia esta pasta.)

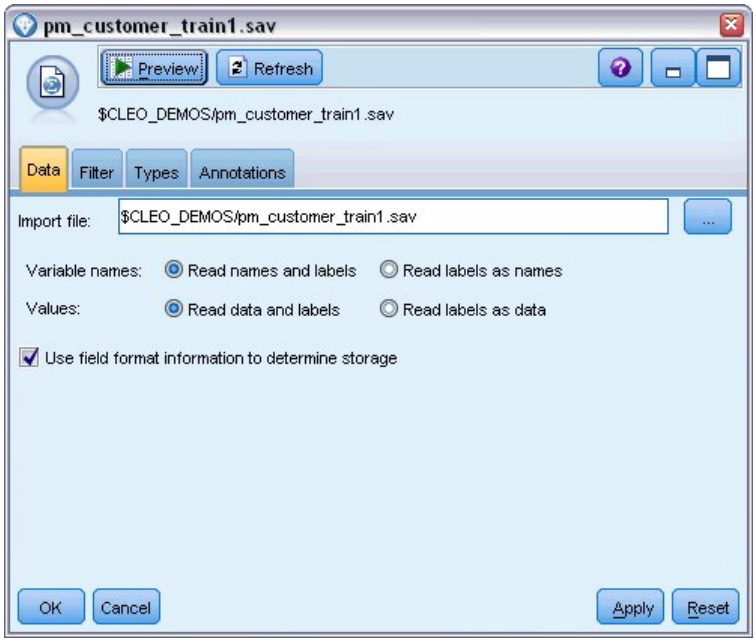


Figura 117. Leitura nos dados

2. Inclua um nó Tipo e selecione *resposta* como o campo de destino (Role = **Destino**). Configure o nível de medição para este campo para **Flag**.

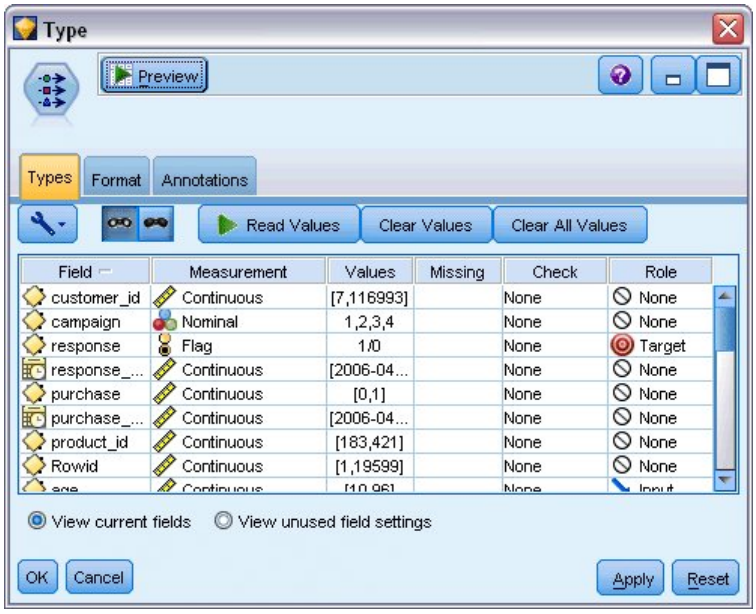


Figura 118. Configurando o nível de medição e a função

3. Configure a função para **Nenhum** para os campos a seguir: *customer_id*, *campanha*, *response_date*, *compra*, *purchase_date*, *product_id*, *Rowid* e *X_random*. Esses campos todos têm usos nos dados mas não serão usados na construção do modelo real.
4. Clique no botão **Valores de leitura** no nó Tipo para ter certeza de que os valores são instanciados.

Embora os dados incluam informações sobre quatro campanhas diferentes, você concentrará a análise em uma campanha por vez. Uma vez que o maior número de registros caem sob a campanha Premium (codificada *campanha = 2* nos dados), você pode usar um nó Select para incluir apenas estes registros no fluxo.

Figura 119. Selecionando registros para uma única campanha

Criando o modelo

1. Anexar um nó da Lista de Decisão no fluxo. Na guia Modelo, configure o **Valor da Destino** para 1 para indicar o resultado que deseja pesquisar. Neste caso, você está procurando por clientes que responderam *Sim* a uma oferta anterior.

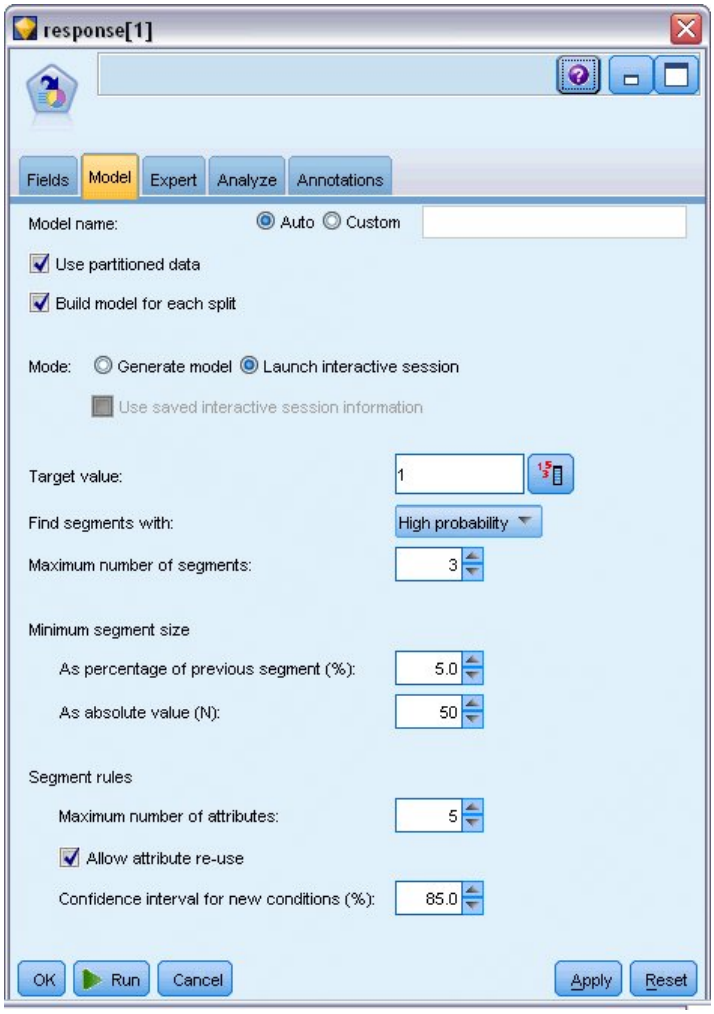


Figura 120. Nó da Lista de Decisão, guia Modelo

2. Selecione **Ativar sessão interativa**.
3. Para manter o modelo simples para finalidades deste exemplo, configure o número máximo de segmentos para 3.
4. Alterar o intervalo de confiança para novas condições para 85%.
5. Na guia Expert, configure o **Mode** para **Expert**.

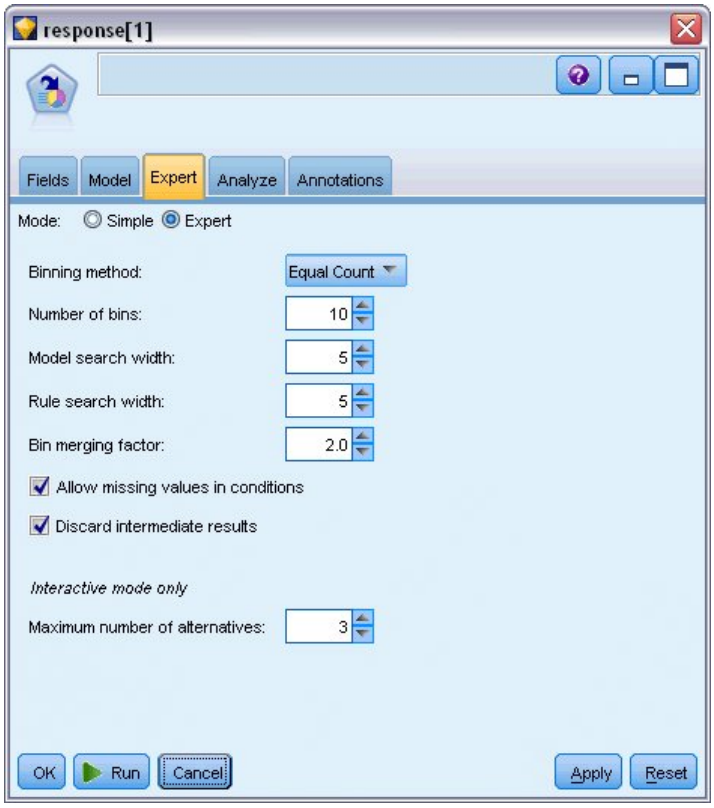


Figura 121. Nó da Lista de Decisão, Guia Expert

6. Aumente o **Número máximo de alternativas** para 3. Esta opção funciona em conjunto com a configuração **Ativar sessão interativa** que você selecionou na guia Modelo.
7. Clique em **Executar** para exibir o visualizador da Lista Interativa.

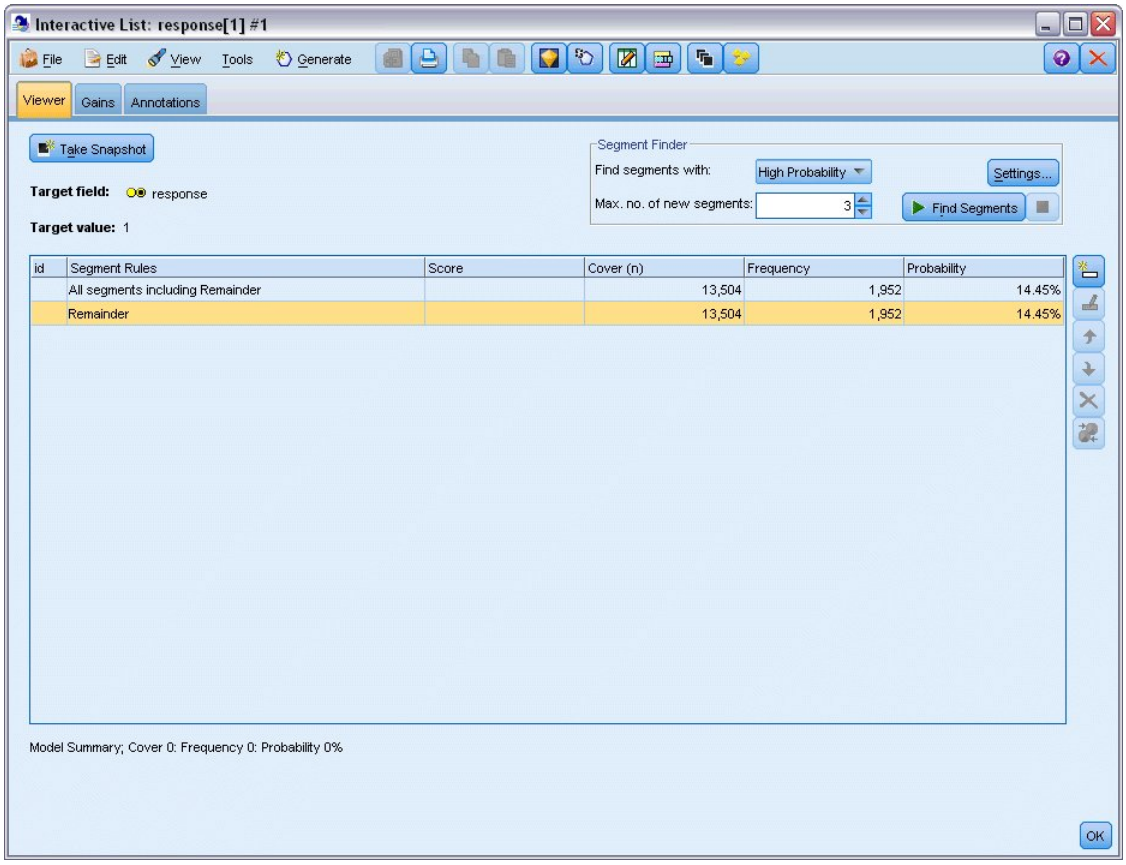


Figura 122. Visualizador de Lista Interativa

Como nenhum segmento ainda foi definido, todos os registros caem sob o restante. De 13.504 registros na amostra 1.952 disseram *Sim*, para uma taxa de ocorrência geral de 14.45%. Você quer melhorar nesta taxa identificando segmentos de clientes mais (ou menos) propensos a dar uma resposta favorável.

8. No visualizador da Lista Interativa, a partir dos menus escolha:

   Ferramentas > **Localiza Segmentos**

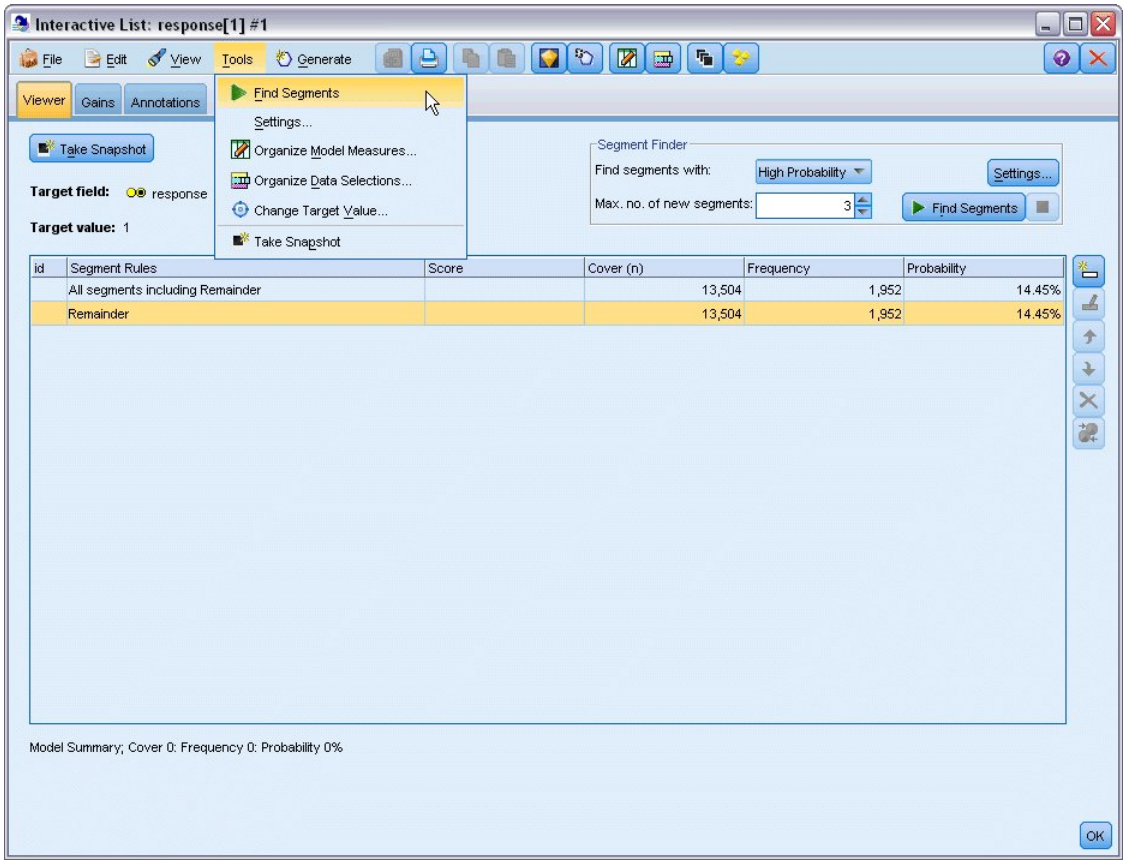


Figura 123. Visualizador de Lista Interativa

Isto executa a tarefa de mineração padrão com base nas configurações que você especificou no nó da Lista de Decisão. A tarefa concluída retorna três modelos alternativos, que estão listados na guia Alternativas da caixa de diálogo Modelo Albums.

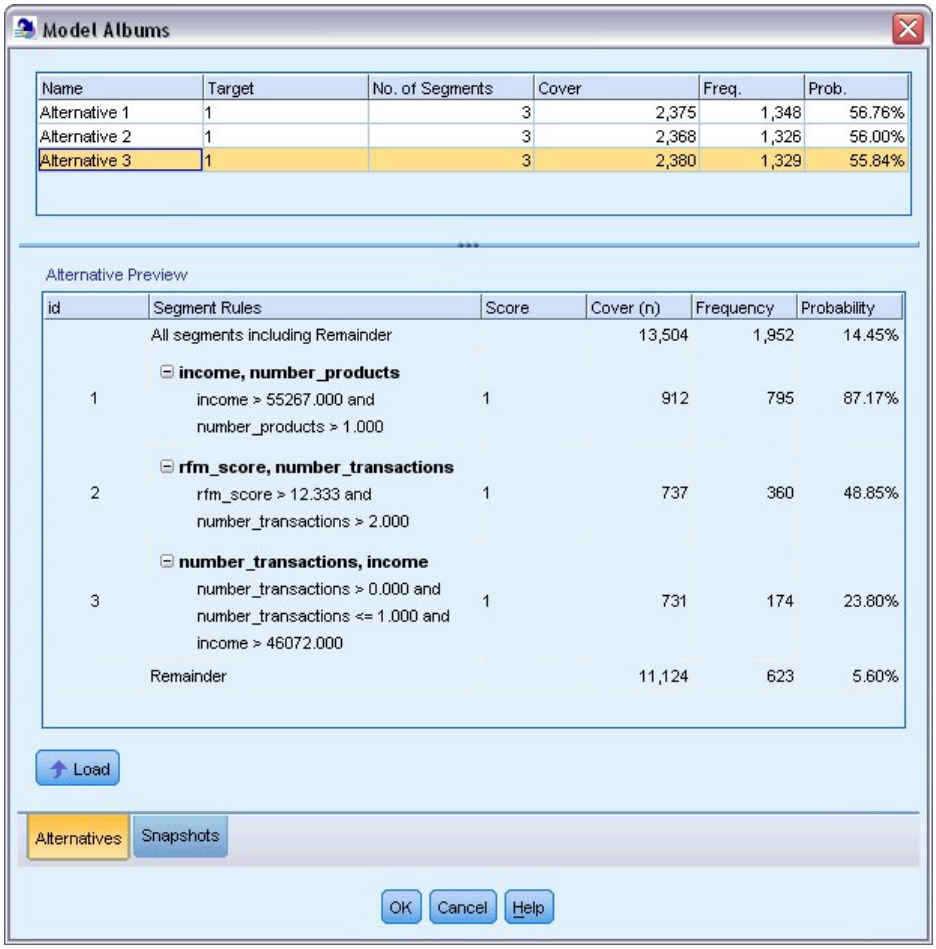


Figura 124. Modelos alternativos disponíveis

9. Selecione a primeira alternativa a partir da lista; seus detalhes são mostrados no painel Previsão Alternativa.

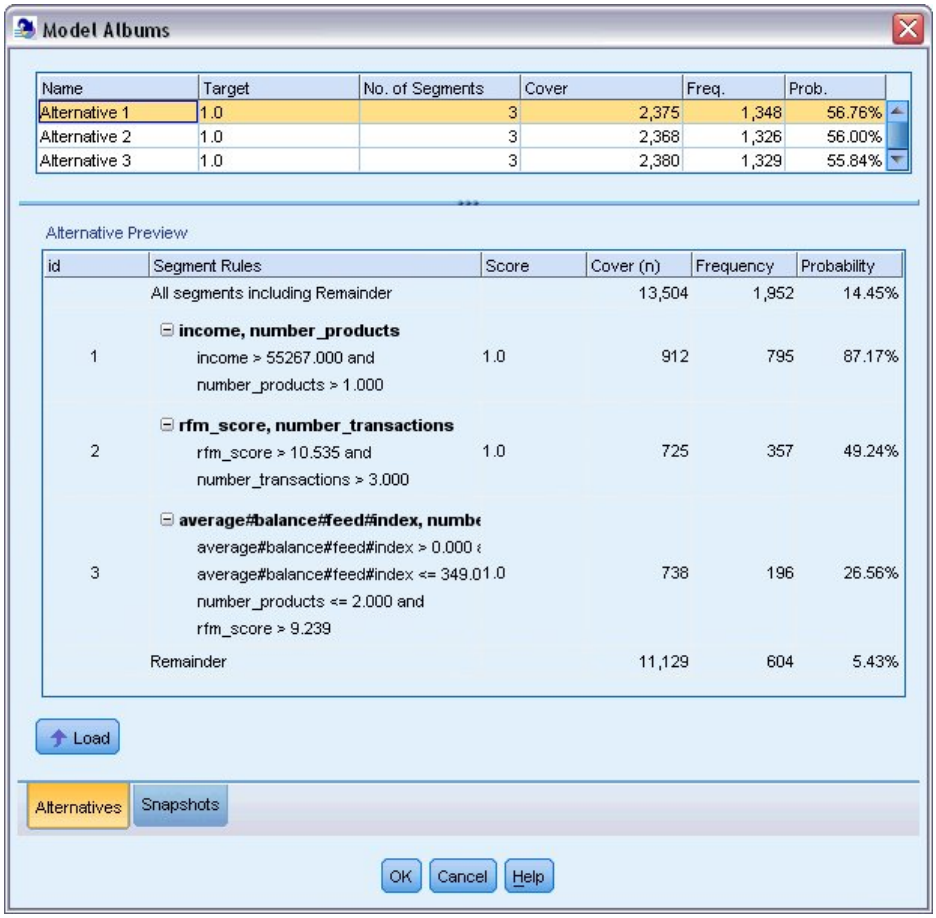


Figura 125. Modelo alternativo selecionado

O Painel Preview Alternativo permite que você navegue rapidamente em qualquer número de alternativas sem alterar o modelo de trabalho, facilando a experiência com diferentes abordagens.

Nota: Para obter um melhor olhar para o modelo, você pode querer maximizar o painel Preview Alternativa dentro do diálogo, como mostrado aqui. Você pode fazer isso arrastando a fronteira do painel.

Usando regras baseadas em preditores, como renda, número de transações por mês e pontuação de RFM, o modelo identifica segmentos com taxas de resposta que são mais altas do que as para a amostra geral. Quando os segmentos são combinados, esse modelo sugere que você poderia melhorar sua taxa de acertos para 56.76% No entanto, o modelo cobre apenas uma pequena parcela da amostra geral, deixando mais de 11.000 registros-com várias centenas de acessos entre eles-para cair sob o restante. Você quer um modelo que capte mais esses acertos enquanto ainda exclui os segmentos de baixo desempenho.

10. Para tentar uma abordagem de modelagem diferente, a partir dos menus escolha:

    Ferramentas > **Configurações**

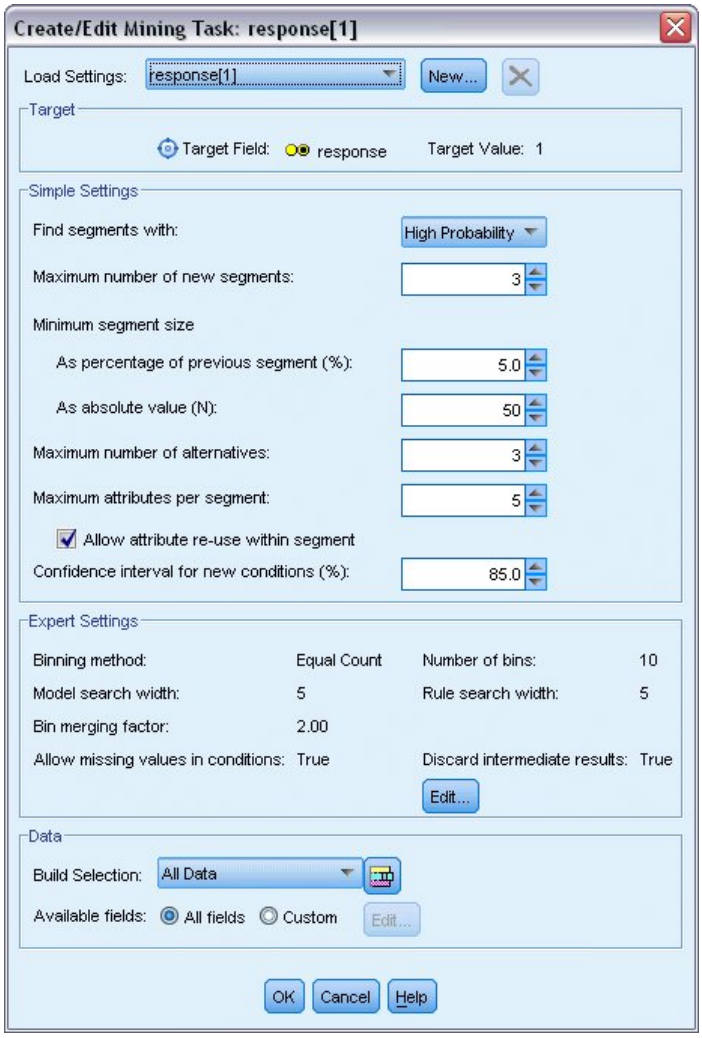


Figura 126. Caixa de Diálogo Criar / Editar Tarefa de Mineração

11. Clique no botão **Novo** (canto superior direito) para criar uma segunda tarefa de mineração, e especifique *Down Search* como o nome da tarefa na caixa de diálogo Nova Configurações.

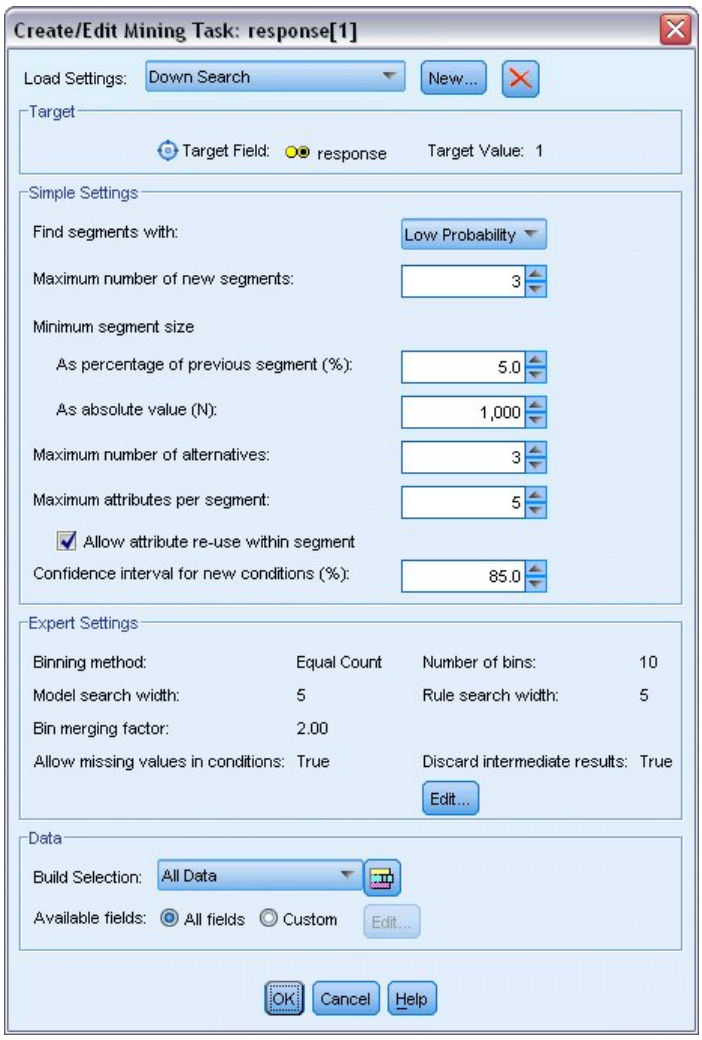


Figura 127. Caixa de Diálogo Criar / Editar Tarefa de Mineração

12. Altere a direção de pesquisa para **Baixa probabilidade** para a tarefa. Isso fará com que o algoritmo pesque por segmentos com as taxas de resposta *mais baixas* em vez das mais altas.
13. Aumente o tamanho do segmento mínimo para 1.000. Clique em **OK** para retornar ao visualizador da Lista Interativa.
14. No Visualizador de Lista Interativa, certise-se de que o painel *Finder de Segmento* esteja exibindo os novos detalhes da tarefa e clique em **Localizar Segmentos**.

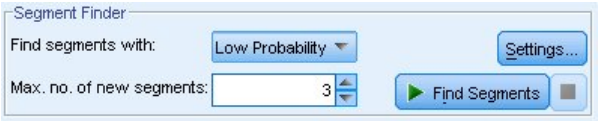


Figura 128. Encontre segmentos em nova tarefa de mineração

A tarefa retorna um novo conjunto de alternativas, que são exibidas na guia Alternativas da caixa de diálogo Modelo Albums e podem ser visualizadas da mesma maneira que os resultados anteriores.

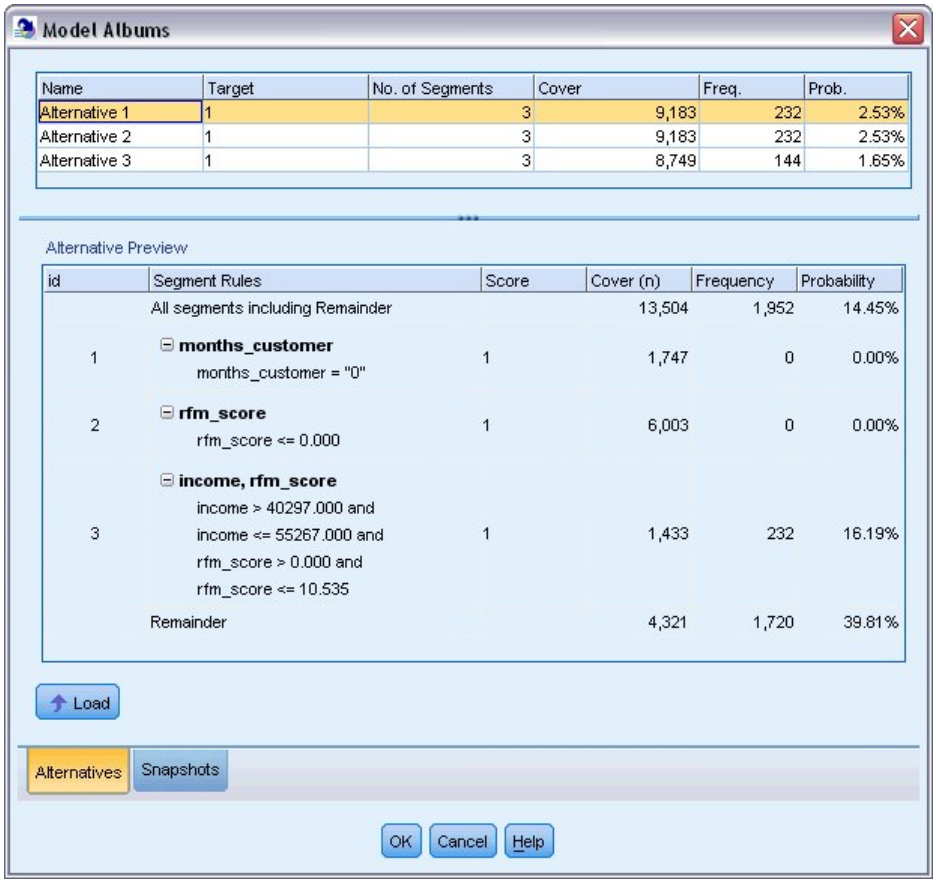


Figura 129. Resultados do modelo de pesquisa

Desta vez cada modelo identifica segmentos com probabilidades de resposta baixa em vez de altos. Observando a primeira alternativa, simplesmente excluir esses segmentos aumentará a taxa de ocorrência para o restante para 39.81%. Isso é mais baixo do que o modelo que você olhou mais cedo mas com cobertura mais alta (significando mais acertos totais).

Ao combinar as duas abordagens-usando uma pesquisa Low Probabilidade para eliminar registros desinteressantes, seguida por uma pesquisa de Alta Probabilidade-você pode ser capaz de melhorar este resultado.

15. Clique em **Carregar** para fazer isso (a primeira alternativa de Down Search) o modelo de trabalho e clique em **OK** para fechar a caixa de diálogo Modelo Albums.

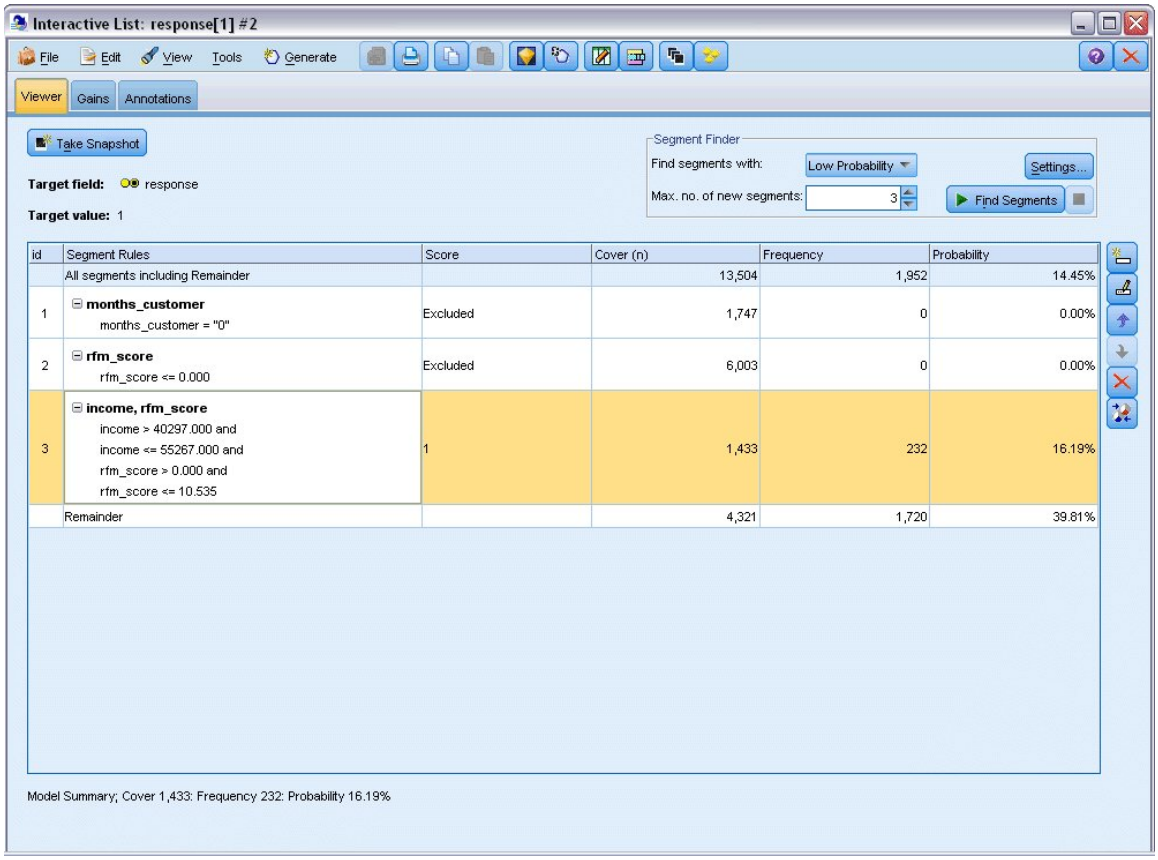


Figura 130. Excluindo um segmento

16. Clique com o botão direito do mouse sobre cada um dos dois primeiros segmentos e selecione **Excluir Segmento**. Juntos, esses segmentos capturam quase 8.000 registros com zero acertos entre eles, por isso faz sentido excluí-los de ofertas futuras. (Segmentos excluídos serão pontuados como nulos para indicar isso.)
17. Clique com o botão direito do mouse sobre o terceiro segmento e selecione **Excluir Segmento**. Em 16.19%, a taxa de ocorrência para esse segmento não é tão diferente da taxa de linha de base de 14.45%, portanto, ela não inclui informações suficientes para justificar mantê-la em vigor

    Nota: Deletar um segmento não é o mesmo que excluindo-o. Excluir um segmento simplesmente muda como ele é pontuado, enquanto o exclui remove-o do modelo inteiramente.

    Tendo excluído os segmentos de menor desempenho, agora é possível pesquisar segmentos de alto desempenho no restante.
18. Clique na linha restante na tabela para selecioná-lo, de modo que a próxima tarefa de mineração se aplique apenas ao restante.

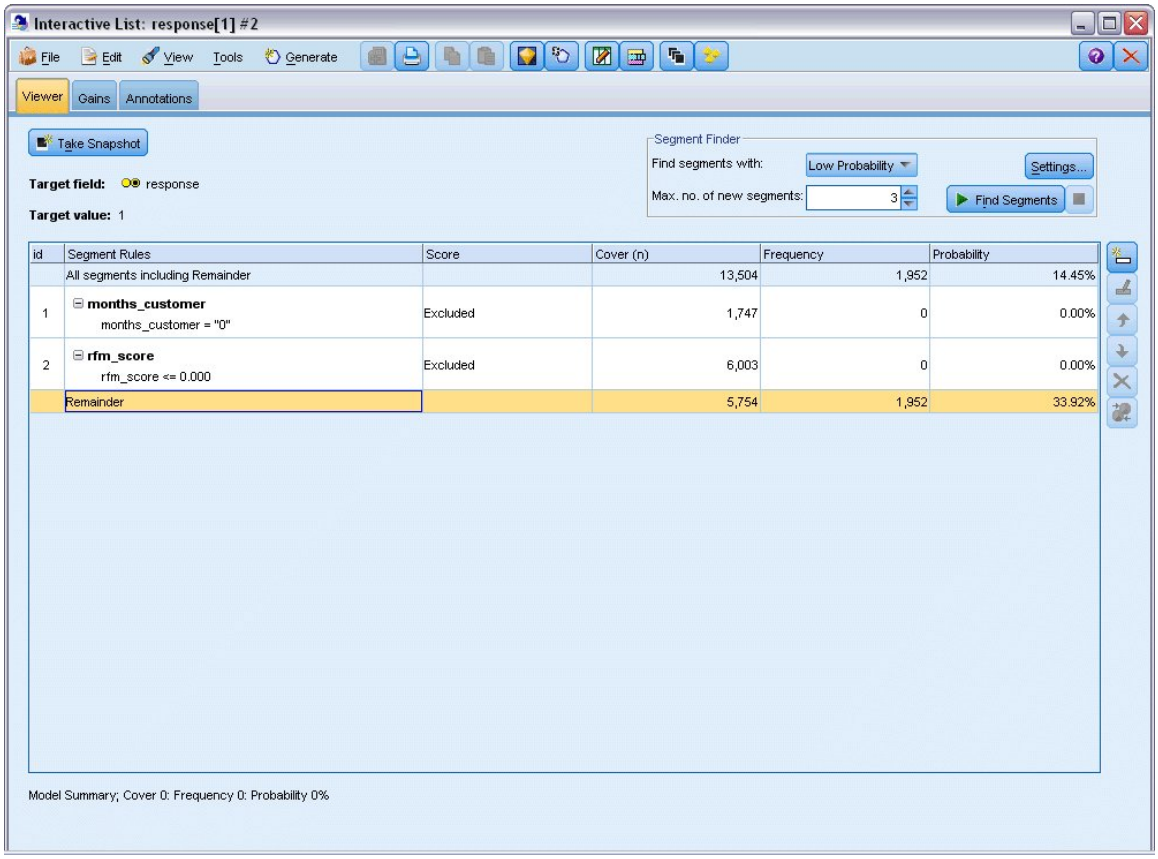


Figura 131. Selecionando um segmento

19. Com o restante selecionado, clique em **Configurações** para reabrir a caixa de diálogo Create / Edit Mining Task.
20. Na parte superior, em **Configurações de carregamento**, selecione a tarefa de mineração padrão: **resposta [1]**.
21. Edite as **Configurações Simples** para aumentar o número de novos segmentos para 5 e o tamanho do segmento mínimo para 500.
22. Clique em **OK** para retornar ao visualizador da Lista Interativa.

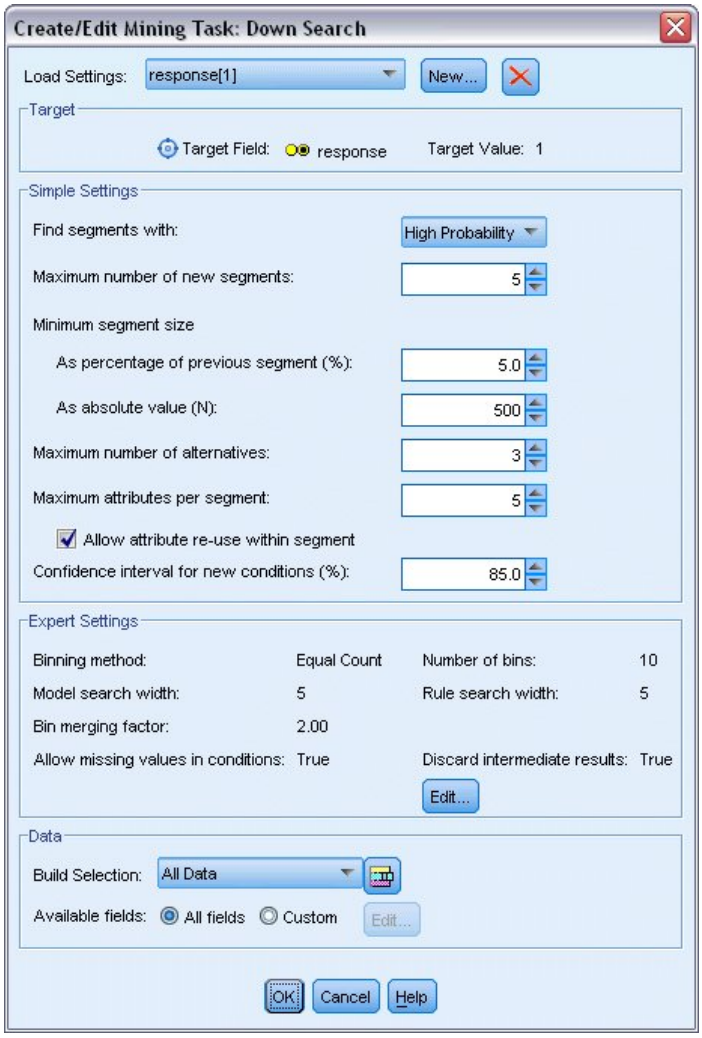


Figura 132. Selecionando a tarefa de mineração padrão

23. Clique em **Encontrar Segmentos**.

    Isso exibe mais um conjunto de modelos alternativos. Ao alimentar os resultados de uma tarefa de mineração em outra, esses modelos mais recentes contêm uma mistura de segmentos de alta e baixa performance. Os segmentos com taxas de resposta baixas são excluídos, o que significa que eles serão pontuados como nulos, enquanto os segmentos incluídos serão marcados como 1. As estatísticas gerais refletem essas exclusões, com o primeiro modelo alternativo mostrando uma taxa de ocorrência de 45.63%, com cobertura mais alta (1.577 ocorrências de 3.456 registros) do que qualquer um dos modelos anteriores.

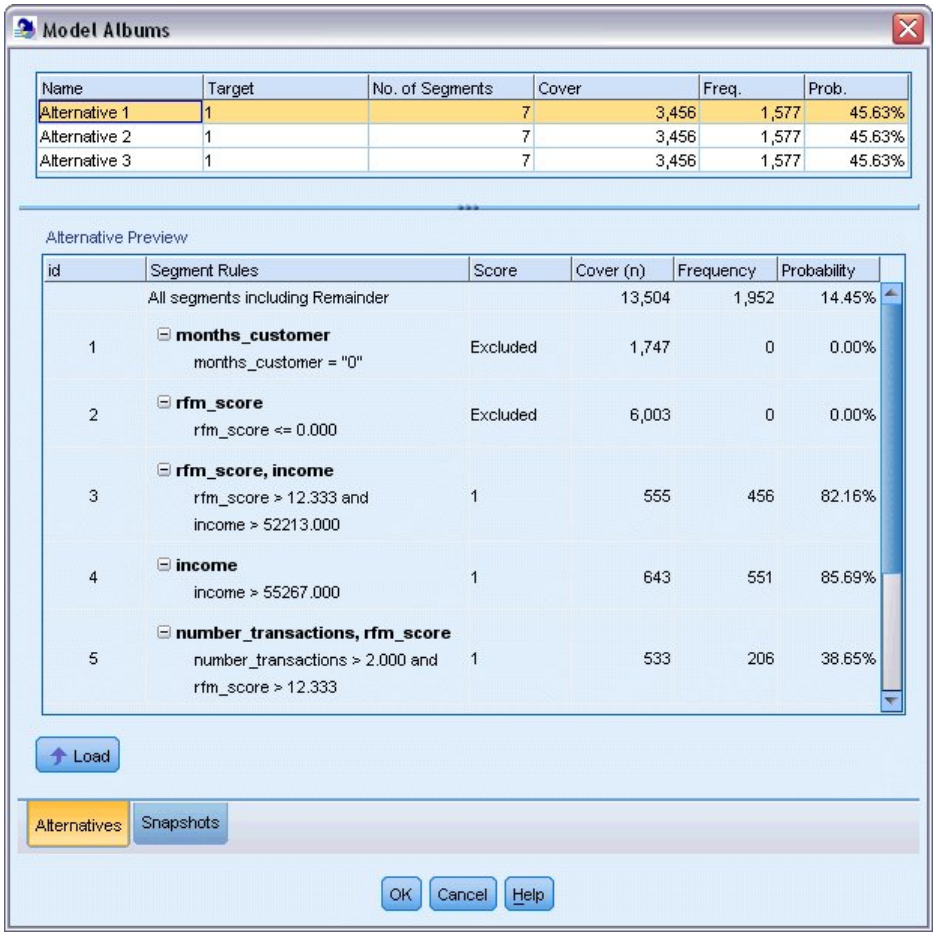


Figura 133. Alternativas para modelo combinado

24. Preview a primeira alternativa e, em seguida, clique em **Carregar** para torná-lo o modelo de trabalho.

Como calcular medidas personalizadas usando o Excel

1. Para ganhar um pouco mais de insight sobre como o modelo realiza em termos práticos, escolha **Organizar Medidas Modelo** a partir do menu Ferramentas.

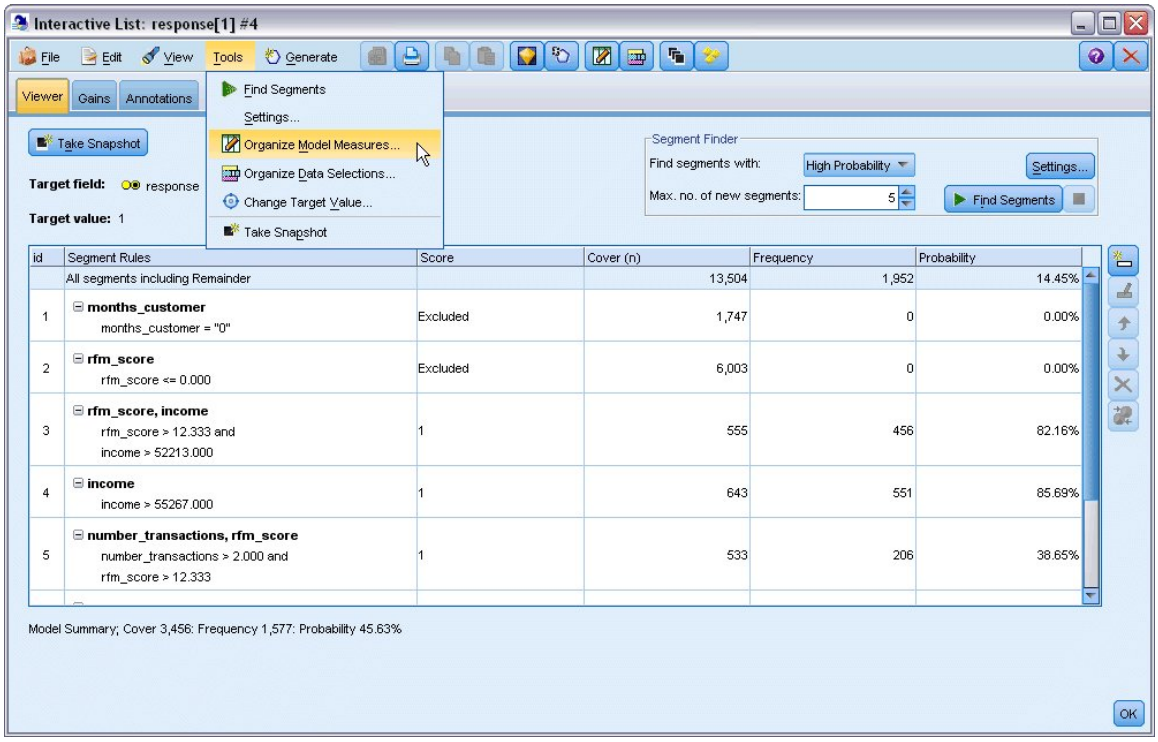


Figura 134. Organizando Medidas de Modelo

A Caixa de diálogo Medidas do Modelo de Organação permite que você escolha as medidas (ou colunas) para mostrar no visualizador da Lista Interativa. Você também pode especificar se as medidas são computadas contra todos os registros ou um subconjunto selecionado, e você pode optar por exibir um gráfico de pizza em vez de um número, quando aplicável.

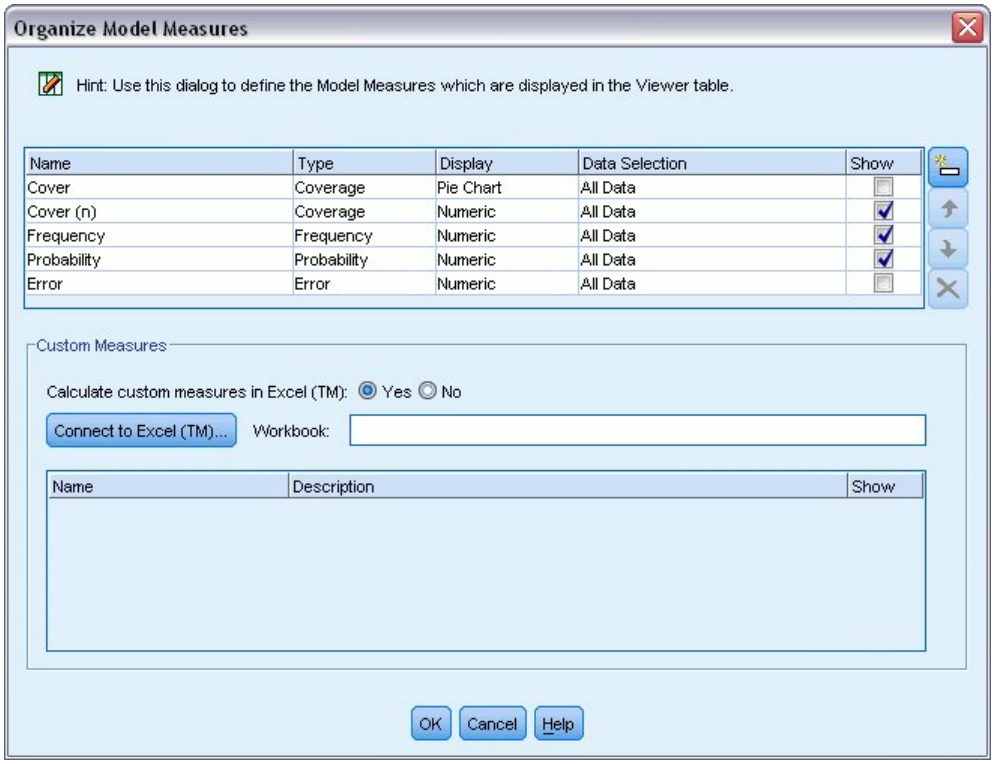


Figura 135. Caixa de diálogo Medidas Modelo de Organação

Além disso, se você tiver o Microsoft Excel instalado, você pode se vincular a um template do Excel que irá calcular as medidas customizadas e adicioná-las ao display interativo.

2. Na caixa de diálogo Organizar Modelo de Medidas, configure **Calcule as medidas customizadas no Excel (TM)** para **Sim**.
3. Clique em **Conectar-se ao Excel (TM)**
4. Selecione a pasta de trabalho *template_profit.xlt* , localizada em *streams* na pasta *Demos* de sua instalação do IBM SPSS Modelador e clique em **Abrir** para ativar a planilha.

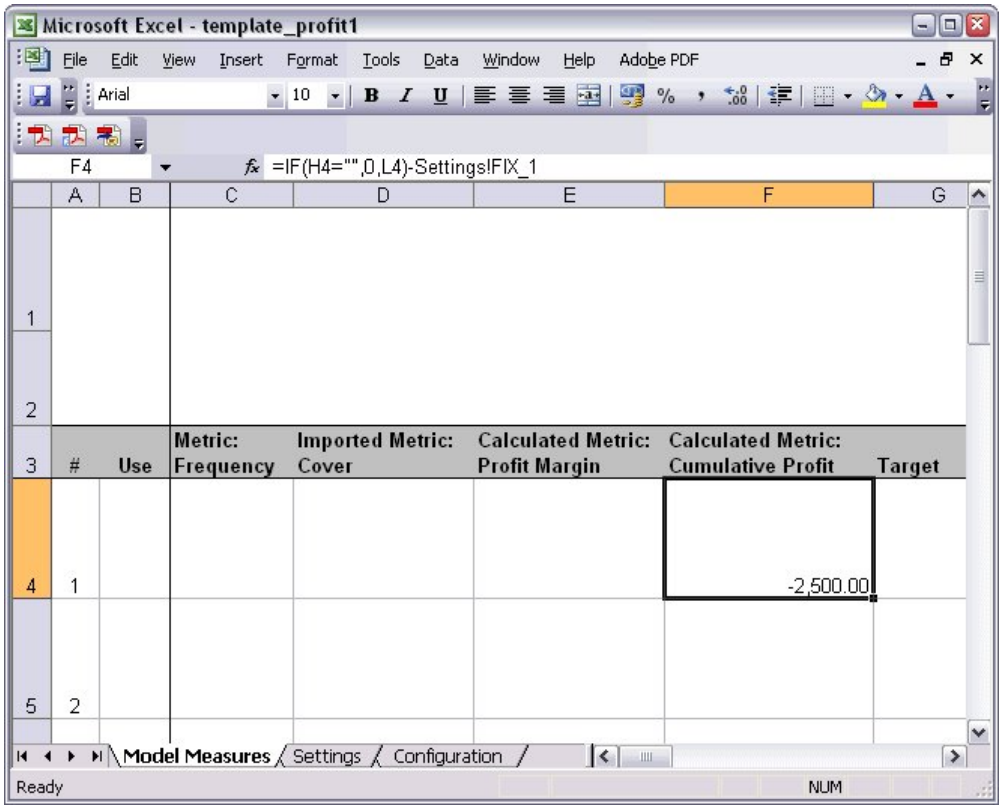


Figura 136. Planilha de Medidas do Modelo Excel

O modelo Excel contém três planilhas:

- **Medidas Modelo** exibe medidas de modelo importadas do modelo e calcula medidas customizadas para exportação de volta para o modelo.
- **Configurações** contém parâmetros a serem usados no cálculo de medidas customizadas.
- **Configuração** define as medidas a serem importadas de e exportadas para o modelo.

As métricas exportadas de volta para o modelo são:

- **Margem De Lucro.** Receita líquida do segmento
- **Lucro acumulado.** Lucro total da campanha

Conforme definido pelas fórmulas a seguir:

```
Profit Margin = Frequency * Revenue per respondent - Cover * Variable cost
Cumulative Profit = Total Profit Margin - Fixed cost
```

Observe que Frequência e Cover são importadas do modelo.

Os parâmetros de custo e receita são especificados pelo usuário na planilha de Configurações.

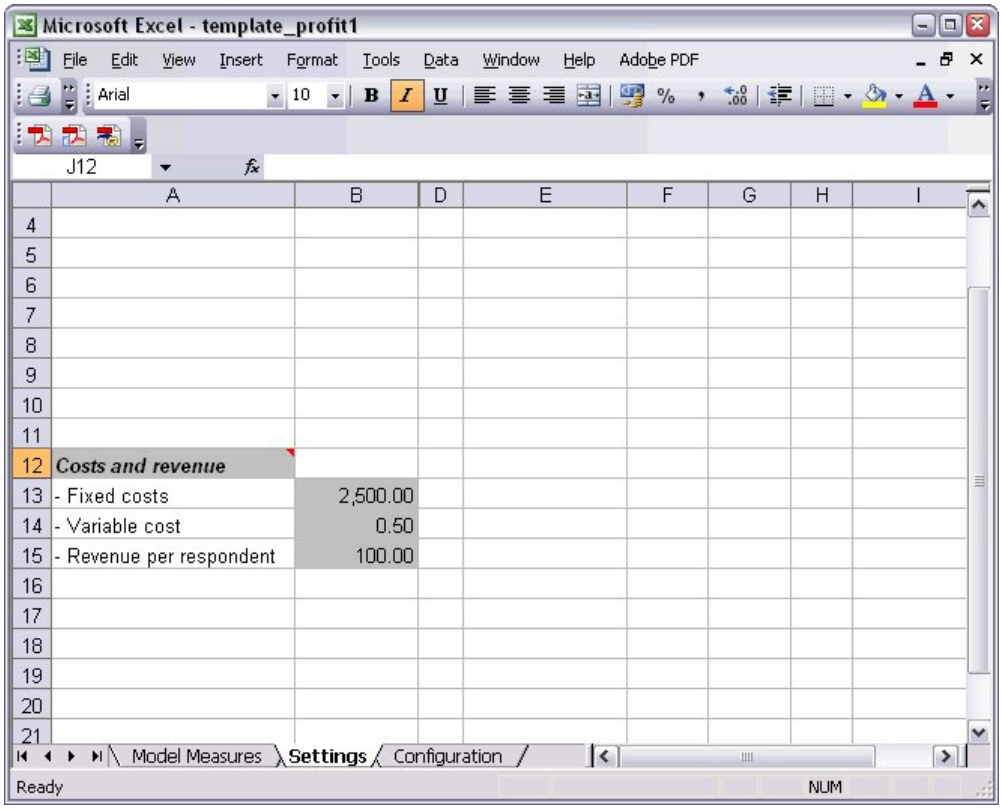


Figura 137. Planilha de Configurações do Excel

Custo fixo é o custo de setup para a campanha, como design e planejamento.

Custo variável é o custo de estender a oferta para cada cliente, como envelopes e estampas.

Receita por respondente é a receita líquida de um cliente que responde pela oferta.

5. Para completar o link de volta para o modelo, use a barra de tarefas do Windows (ou pressione Alt + Tab) para navegar de volta para o visualizador da Lista Interativa.

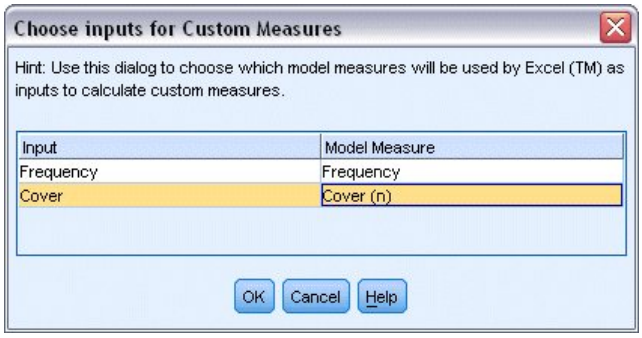


Figura 138. Escolhendo entradas para medidas personalizadas

A Caixa de diálogo Escolher Entradas para Medidas Personalizadas é exibida, permitindo mapear entradas do modelo para parâmetros específicos definidos no template. A coluna esquerda lista as medidas disponíveis, e a coluna da direita mapeia estes para planilha de parâmetros conforme definido na planilha de Configuração.

6. Na coluna **Medidas Modelo** , selecione **Frequência** e **Cover (n)** contra as respectivas entradas e clique em **OK**.

   Nesse caso, os nomes de parâmetros no template-Frequency e Cover (n)-acontecem para combinar com as entradas, mas nomes diferentes também poderiam ser usados.

7. Clique em **OK** na caixa de diálogo Organizar Modelo de Medidas para atualizar o visualizador da Lista Interativa.

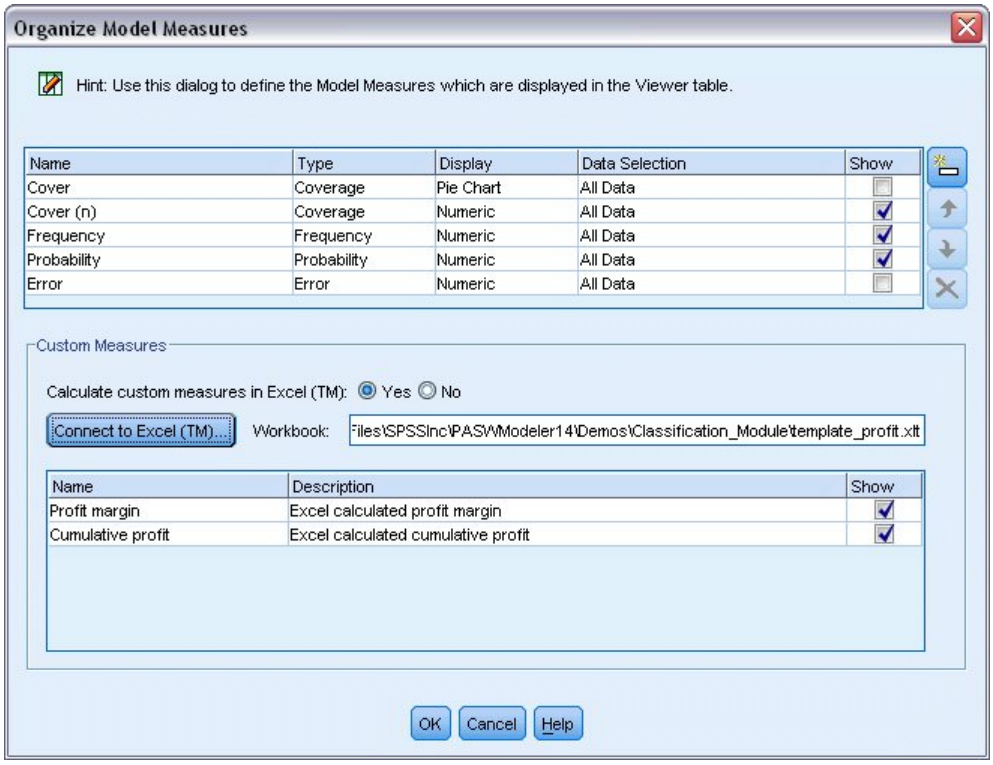


Figura 139. Organizar caixa de diálogo Medidas Modelo mostrando medidas personalizadas do Excel

As novas medidas agora são adicionadas como novas colunas na janela e serão recalculadas a cada vez que o modelo for atualizado.

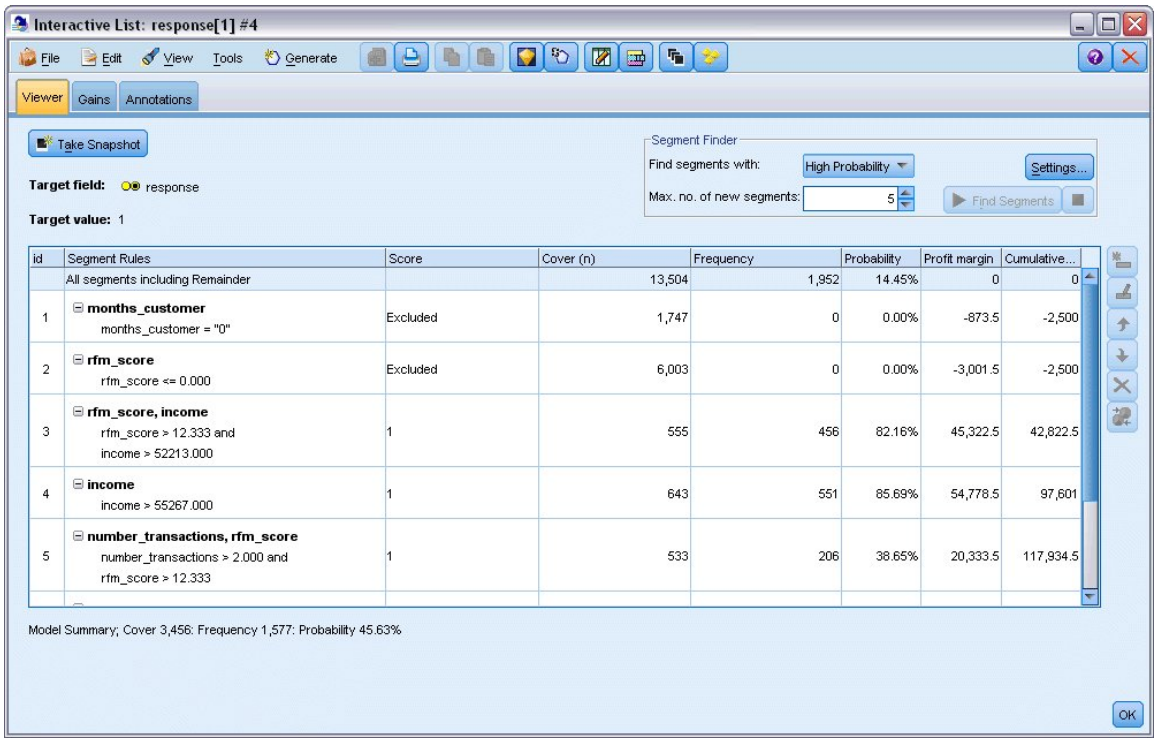


Figura 140. Medidas customizadas do Excel exibidas no visualizador de Lista Interativa

Ao editar o modelo Excel, qualquer número de medidas customizadas pode ser criado.

Modificando o template do Excel

Apesar de IBM SPSS Modelador ser fornecido com um template padrão do Excel para usar com o visualizador de Lista Interativa, você pode querer alterar as configurações ou adicionar a sua própria. Por exemplo, os custos no gabarito podem estar incorretos para a sua organização e necessidade de alteração.

Nota: Se você fizer modificar um modelo existente, ou criar você mesmo, lembre-se de salvar o arquivo com um sufixo Excel 2003 *.xlt* .

Para modificar o modelo padrão com novos detalhes de custo e receita e atualizar o visualizador da Lista Interativa com os novos números:

1. No visualizador da Lista Interativa, escolha **Organizar Medidas Modelo** a partir do menu Ferramentas.
2. Na caixa de diálogo Organizar Modelo de Medidas, clique em **Conectar-se ao Excel™.**
3. Selecione a planilha *template_profit.xlt* e clique em **Abrir** para ativar a planilha.
4. Selecione a planilha Configurações.
5. Edite **Custos fixos** para 3,250.00e **Renda por respondente** para 150.00.

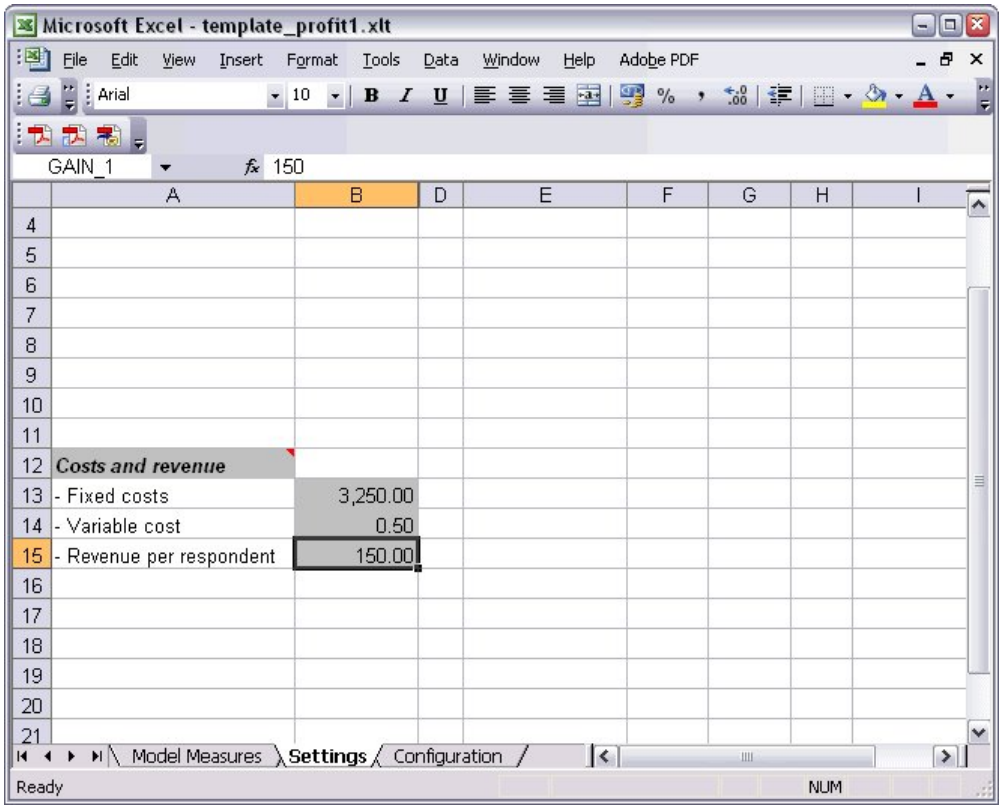


Figura 141. Valores modificados na planilha de Configurações do Excel

6. Salve o modelo modificado com um filename exclusivo e relevante. Certifica-se de possuir uma extensão Excel 2003 *.xlt* .

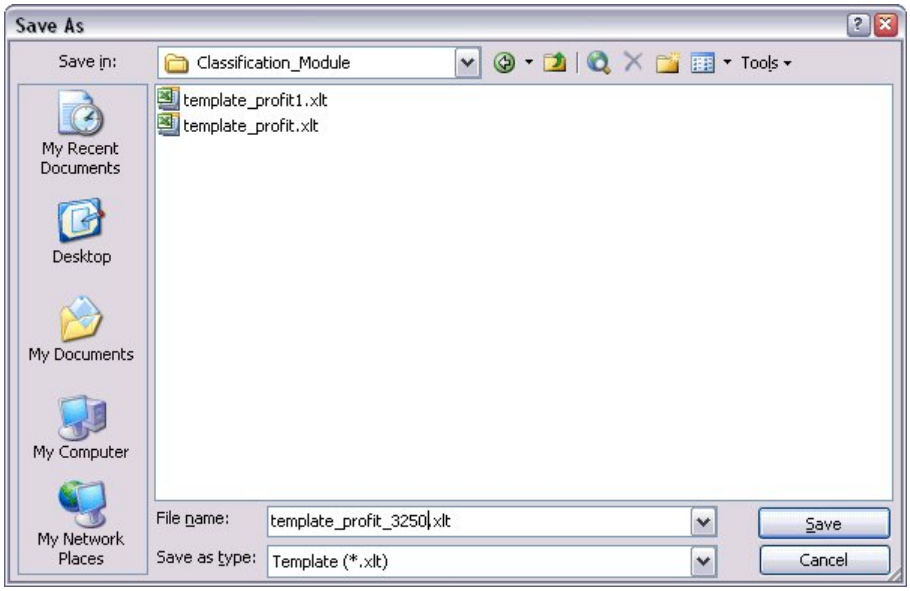


Figura 142. Salvando modelo de Excel modificado

7. Use a barra de tarefas do Windows (ou pressione Alt + Tab) para navegar de volta para o visualizador da Lista Interativa.

   Na caixa de diálogo Escolher Entradas para Medidas Customizadas, selecione as medidas que deseja exibir e clique em **OK**.

8. Na caixa de diálogo de Medidas do Modelo Organize, clique em **OK** para atualizar o visualizador da Lista Interativa.

Obviamente, este exemplo tem mostrado apenas uma maneira simples de modificar o template do Excel; você pode fazer novas alterações que puxam dados de, e passar dados para, o visualizador de Lista Interativa ou trabalhar dentro do Excel para produzir outras saídas, como gráficos.

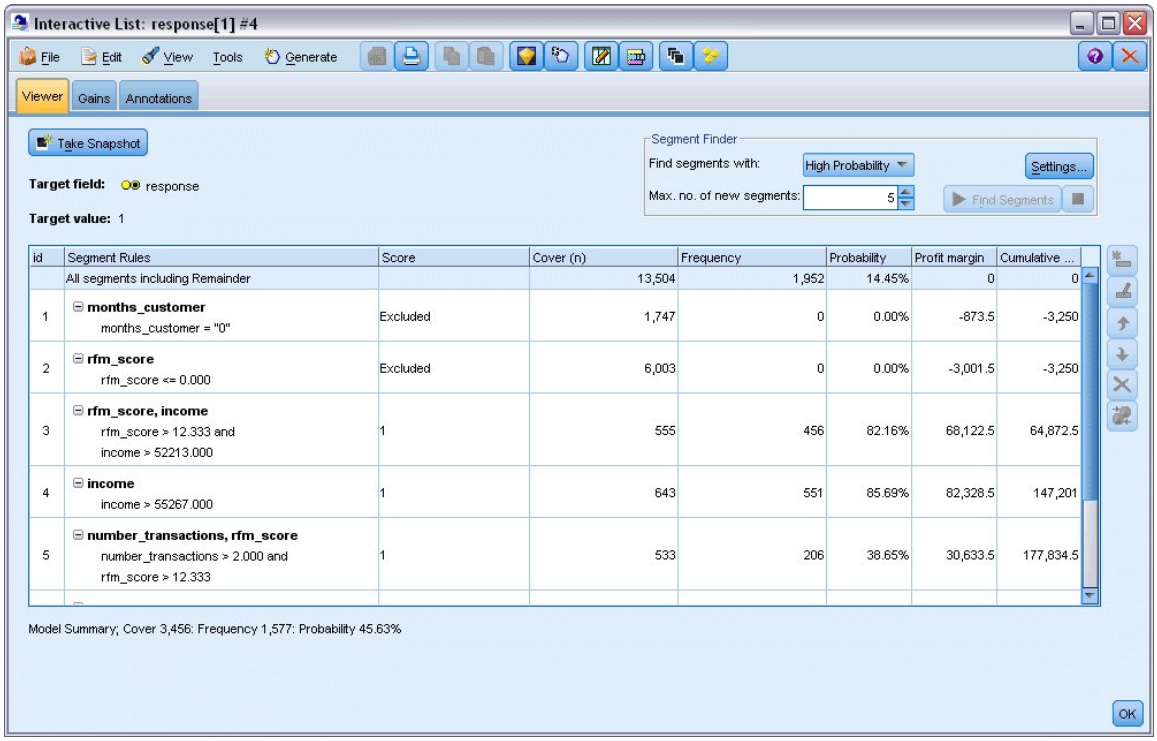


Figura 143. Medidas personalizadas modificadas do Excel exibidas no visualizador de Lista Interativa

Salvando os Resultados

Para salvar um modelo para uso posterior durante a sua sessão interativa, você pode tirar um instantâneo do modelo, que será listado na guia Snapshots. Você pode retornar a qualquer instantâneo salvo a qualquer momento durante a sessão interativa.

Continuando desta maneira, você pode experimentar tarefas adicionais de mineração para pesquisar por segmentos adicionais. Você também pode editar segmentos existentes, inserir segmentos personalizados com base em suas próprias regras de negócio, criar seleções de dados para otimizar o modelo para grupos específicos e customizar o modelo de uma série de outras maneiras. Por fim, você pode incluir explicitamente ou excluir cada segmento como apropriado para especificar como cada um será pontuado.

Quando estiver satisfeito com seus resultados, você pode usar o menu Gerar para gerar um modelo que pode ser adicionado a fluxos ou implantado para fins de pontuação.

Alternativamente, para salvar o estado atual de sua sessão interativa por mais um dia, escolha **Atualizar o Nó de Modelagem** no menu Arquivo. Isso atualizará o nó de modelagem da Lista de Decisão com as configurações atuais, incluindo tarefas de mineração, snapshots de modelo, seleções de dados e medidas customizadas. Na próxima vez que você executar o fluxo, apenas certifique-se de que **Usar informações de sessão salva** é selecionado no nó de modelagem da Lista de Decisão para restaurar a sessão para o seu estado atual.

Capítulo 12. Classificando Os Clientes De Telecomunicações (Regressão Logística Multinomial)

Regressão logística é uma técnica estatística para classificar registros com base em valores de campos de entrada. Ela é semelhante a uma regressão linear, mas usa um campo de destino categórico ao invés de um campo numérico.

Por exemplo, suponha que um provedor de Telecomunicações tenha segmentado sua base de clientes por padrões de uso de serviço, categorizando os clientes em quatro grupos. Se os dados demográficos puderem ser usados para prever a associação ao grupo, será possível customizar ofertas para clientes em potencial individuais.

Este exemplo usa o fluxo denominado *telco_custcat.str*, que faz referência ao arquivo de dados denominado *telco.sav* Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *telco_custcat.str* está no diretório *streams* .

O exemplo se concentra no uso de dados demográficos para prever os padrões de uso. O campo de destino *custcat* tem quatro valores possíveis que correspondem aos quatro grupos de clientes, como segue:

Valor	**Rótulo**
1	Serviço Básico
2	E-Service
3	Serviço Plus
4	Serviço total

Como o destino tem várias categorias, um modelo multinomial é usado. No caso de um destino com duas categorias distintas, como sim/não, verdadeiro/falso ou rotativo/não rotativo, um modelo binomial poderia ser criado em seu lugar. Consulte o tópico Capítulo 13, “Churn de Telecomunicações (Regressão Logística Binomial)”, na página 129 para obter informações adicionais.

Construindo o Fluxo

1. Inclua um nó de origem do Arquivo de Estatísticas apontando para *telco.sav* na pasta *Demos* .

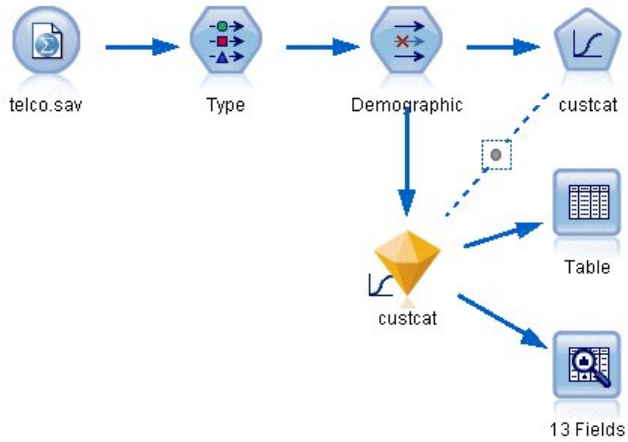


Figura 144. Fluxo de amostra para classificar os clientes usando regressão logística multinomial

a. Adiciize um nó Tipo e clique em **Valores de leitura**, certificando-se de que todos os níveis de medição estão configurados corretamente. Por exemplo, a maioria dos campos com valores 0 e 1 pode ser considerada como sinalizadores.

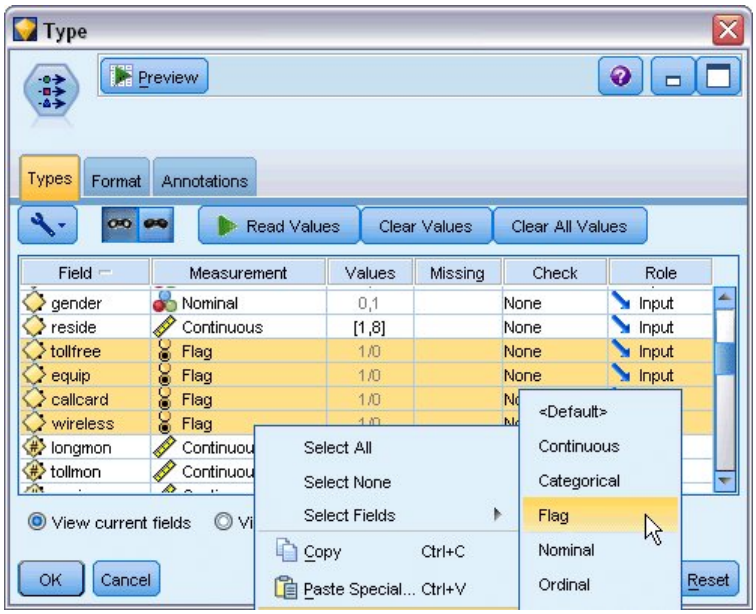


Figura 145. Configurando o nível de medição para vários campos

Sugestão: Para alterar propriedades para vários campos com valores semelhantes (como 0/ 1), clique no cabeçalho da coluna *Valores* para classificar campos por valor e, em seguida, mantenha a tecla shift ao utilizar as teclas do mouse ou seta para selecionar todos os campos que deseja alterar. Em seguida, é possível clicar com o botão direito do mouse na seleção para alterar o nível de medição ou outros atributos dos campos selecionados.

Observe que o *gênero* é considerado mais corretamente como um campo com um conjunto de dois valores, em vez de um sinalizador, portanto, deixe seu valor de medição como **Nominal**.

b. Configure a função para o campo *custcat* como **Destino**. Todos os outros campos devem ter seu papel configurado como **Entrada**.

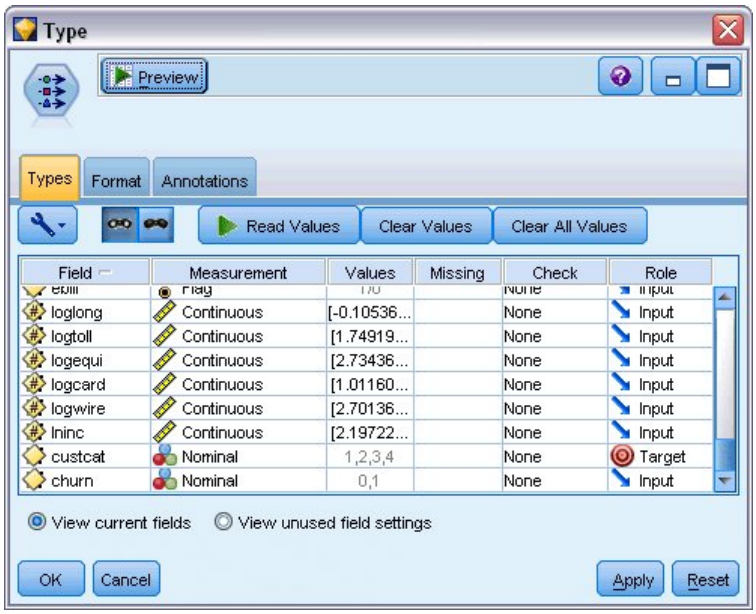


Figura 146. Configurando função de campo

Uma vez que este exemplo se concentra em demografia, use um nó Filtro para incluir apenas os campos relevantes (*região*, *idade*, *marital*, *endereço*, *renda*, *ed*, *empregar*, *aposentadoria*, *gênero*, *residir*, e *custcat*). Outros campos podem ser excluídos com a finalidade desta análise.

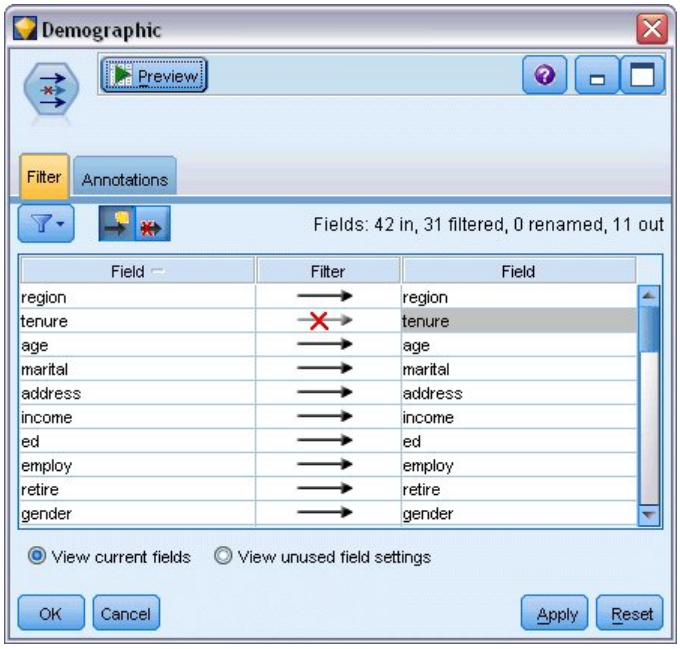


Figura 147. Filtragem em campos demográficos

(Alternativamente, você poderia alterar a função para **Nenhum** para esses campos em vez de excluí-los, ou selecionar os campos que deseja utilizar no nó de modelagem.)

2. No nó Logístico, clique na guia **Modelo** e selecione o método **Stepwise** . Selecione **Multinomial**, **Principais Efeitos**, e **Include constante na equação** também.

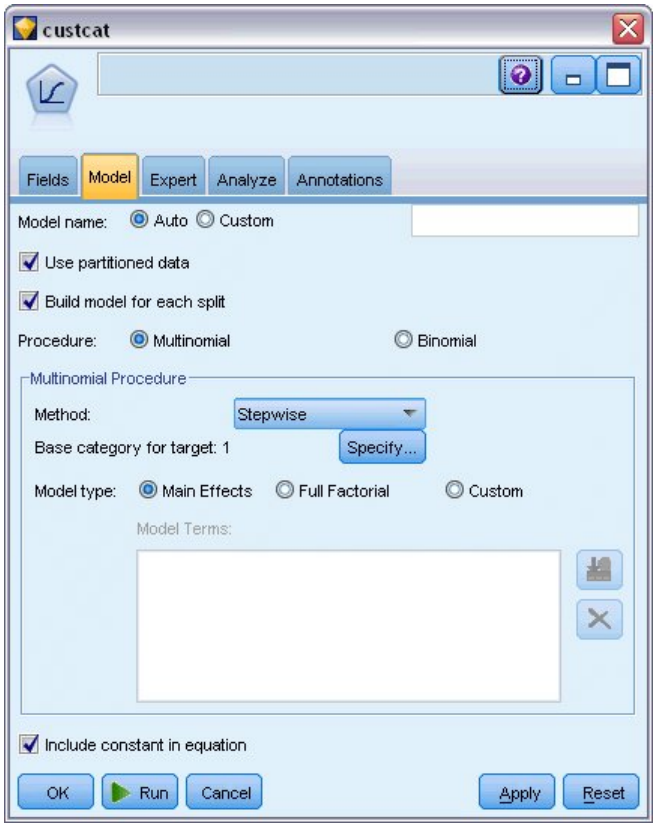


Figura 148. Escolhendo opções de modelo

Deixe a categoria de Base para destino como 1. O modelo vai comparar outros clientes com aqueles que assinam o Serviço Básico.

3. Na guia Expert, selecione o modo **Expert** , selecione **Output**e, na caixa de diálogo Advanced Output, selecione **Tabela de Classificação**.

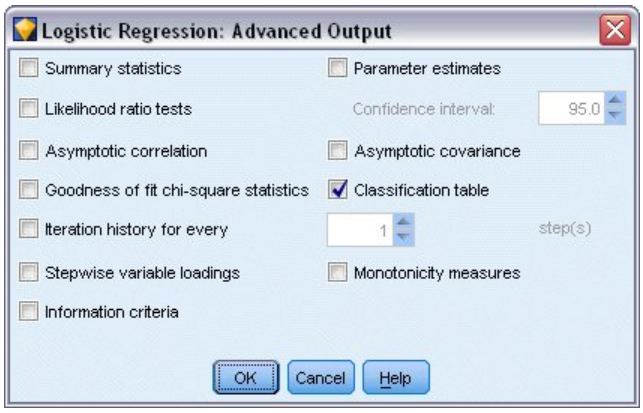


Figura 149. Escolhendo opções de saída

Procurando o Modelo

1. Execute o nó para gerar o modelo, que é adicionado à paleta de Models no canto superior direito. Para visualizar seus detalhes, clique com o botão direito do mouse sobre o nó do modelo gerado e escolha **Browse**.

A guia modelo exibe as equações usadas para designar registros a cada categoria do campo de destino. Há quatro categorias possíveis, uma delas é a categoria de base para a qual não são mostrados detalhes de equação. Os detalhes são mostrados para as três equações restantes, em que a categoria 3 representa o Plus Service, e assim por diante.

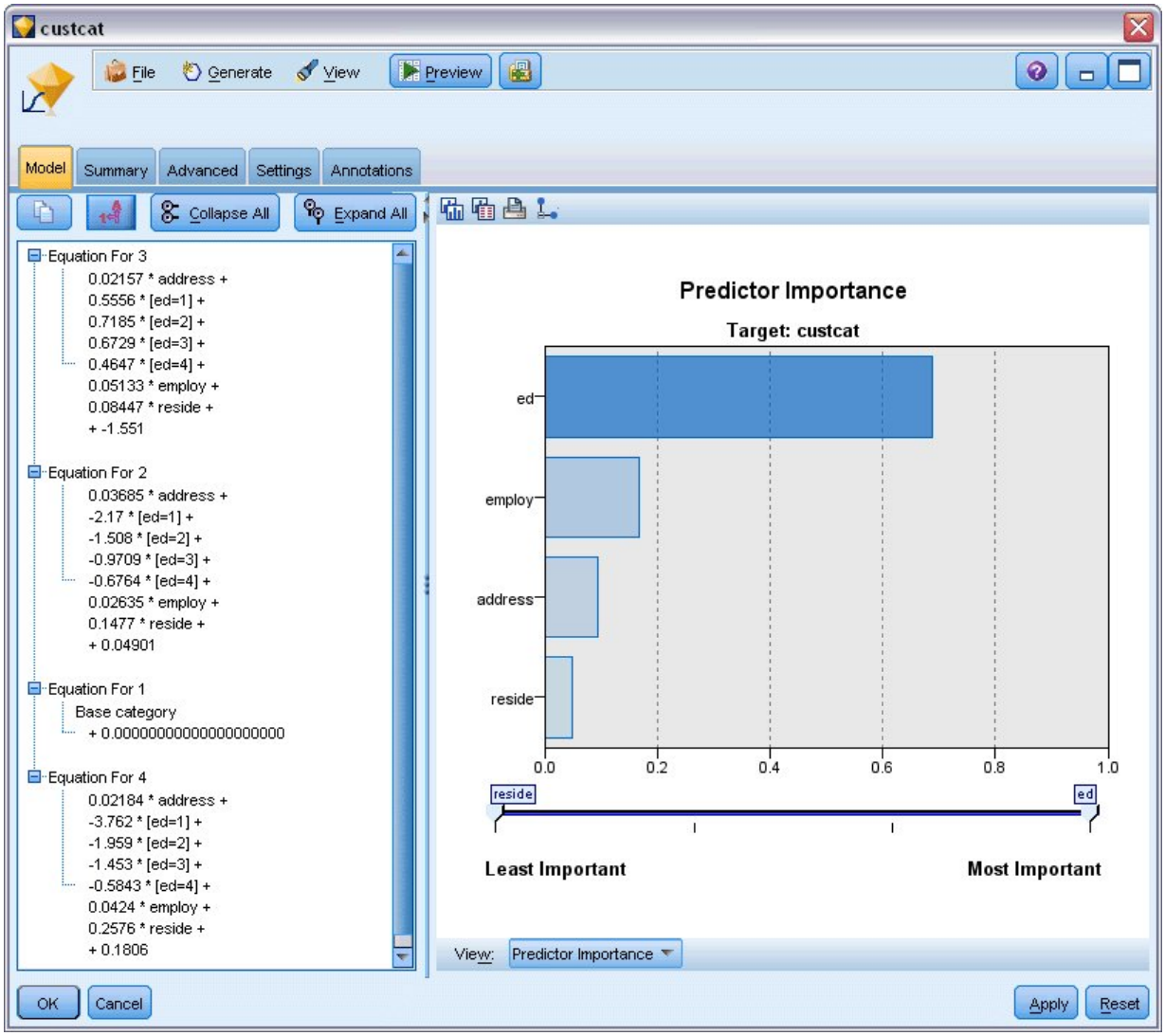


Figura 150. Procurando os resultados do modelo

A guia Resumo mostra (entre outras coisas) o destino e entradas (campos do preditor) utilizados pelo modelo. Observe que estes são os campos que foram realmente escolhidos com base no método Stepwise, não a lista completa enviada para consideração.

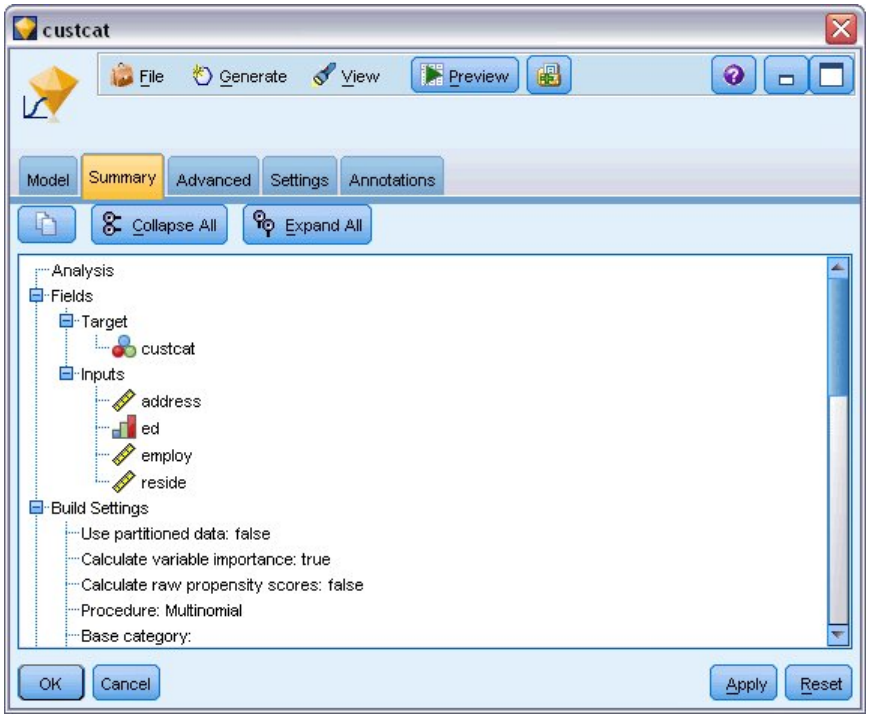


Figura 151. Sumário modelo mostrando campos de destino e entrada

Os itens mostrados na guia Avançado dependem das opções selecionadas na caixa de diálogo de Saída Avançada no nó da modelagem.

Um item que sempre é mostrado é o Sumário de Processamento de Caso, que mostra a porcentagem de registros que cai em cada categoria do campo de destino. Isso dá um modelo nulo para usar como base de comparação.

Sem construir um modelo que utilizasse preditores, o seu melhor palpite seria atribuir todos os clientes ao grupo mais comum, que é o do serviço Plus.

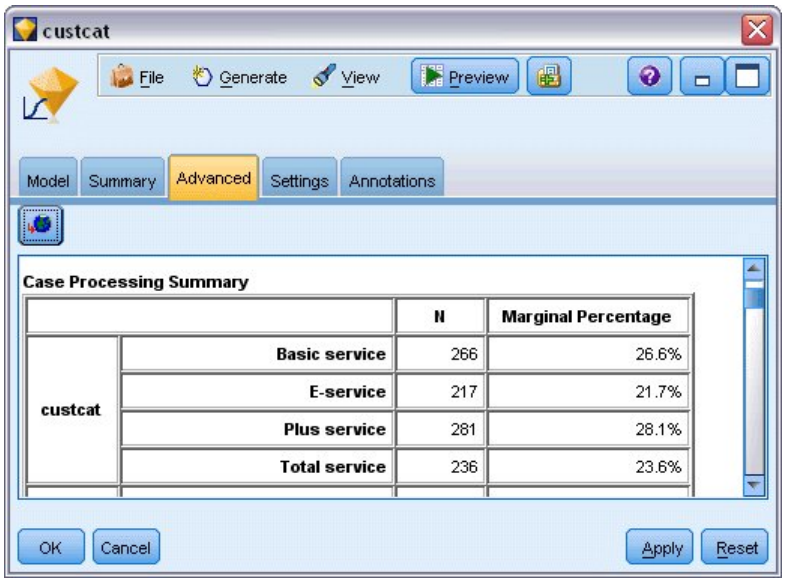


Figura 152. Sumarização do processamento de caso

Com base nos dados de treinamento, se você designasse todos os clientes ao modelo nulo, estaria correto 281/1000 = 28.1% do tempo. A guia Avançado contém mais informações que possibilitam examinar as previsões do modelo. Você pode então comparar as previsões com os resultados do modelo nulo para ver o quão bem o modelo funciona com seus dados.

Na parte inferior da guia Avançado, a tabela Classificação mostra os resultados para seu modelo, que está correto em 39.9% do tempo

Em particular, o seu modelo se sobressai na identificação dos clientes do Total Service (categoria 4) mas faz um trabalho muito pobre de identificação dos clientes do E-service (categoria 2). Se você deseja uma melhor precisão para os clientes na categoria 2, pode ser necessário encontrar outro preditor para identificá-los.

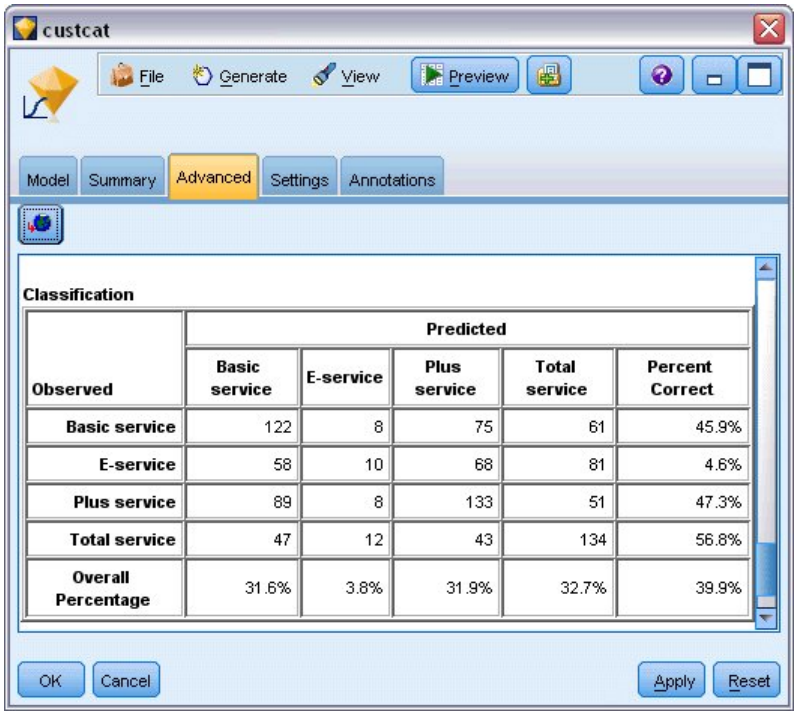


Figura 153. Tabela de classificação

Dependendo do que você quer prever, o modelo pode estar perfeitamente adequado para as suas necessidades. Por exemplo, se você não está preocupado em identificar clientes na categoria 2, o modelo pode ser preciso o suficiente para você. Este pode ser o caso em que o E-service é um líder deficitária que traz pouco lucro.

Se, por exemplo, o seu maior retorno sobre investimento vier de clientes que se enquadram na categoria 3 ou 4, o modelo pode lhe dar a informação de que precisa.

Para avaliar o quão bem o modelo realmente se encaixa nos dados, vários diagnósticos estão disponíveis na caixa de diálogo Advanced Output quando você está construindo o modelo. Explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no *IBM SPSS Modelador Algorithms Guide*, disponível a partir do diretório *Documentação* do disco de instalação.

Observe também que esses resultados são baseados apenas nos dados de treinamento. Para avaliar o quão bem o modelo generaliza para outros dados no mundo real, é possível usar um nó de partição para conter um subconjunto de registros para fins de teste e validação.

Capítulo 13. Churn de Telecomunicações (Regressão Logística Binomial)

Regressão logística é uma técnica estatística para classificar registros com base em valores de campos de entrada. Ela é semelhante a uma regressão linear, mas usa um campo de destino categórico ao invés de um campo numérico.

Este exemplo usa o fluxo denominado *telco_churn.str*, que faz referência ao arquivo de dados denominado *telco.sav* Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *telco_churn.str* está no diretório *streams* .

Por exemplo, suponha que um provedor de telecomunicações esteja preocupado com o número de clientes que está perdendo para os concorrentes. Se os dados de uso do serviço puderem ser usados para prever quais clientes estão sujeitos a transferência para outro provedor, as ofertas podem ser personalizadas para reter o maior número possível de clientes.

Este exemplo se concentra no uso de dados de uso para prever a perda de clientes (rotatividade). Como o destino tem duas categorias distintas, um modelo binomial é usado. No caso de um alvo com várias categorias, um modelo multinomial pode ser criado em seu lugar. Consulte o tópico Capítulo 12, “Classificando Os Clientes De Telecomunicações (Regressão Logística Multinomial)”, na página 121 para obter informações adicionais.

Construindo o Fluxo

1. Inclua um nó de origem do Arquivo de Estatísticas apontando para *telco.sav* na pasta *Demos* .

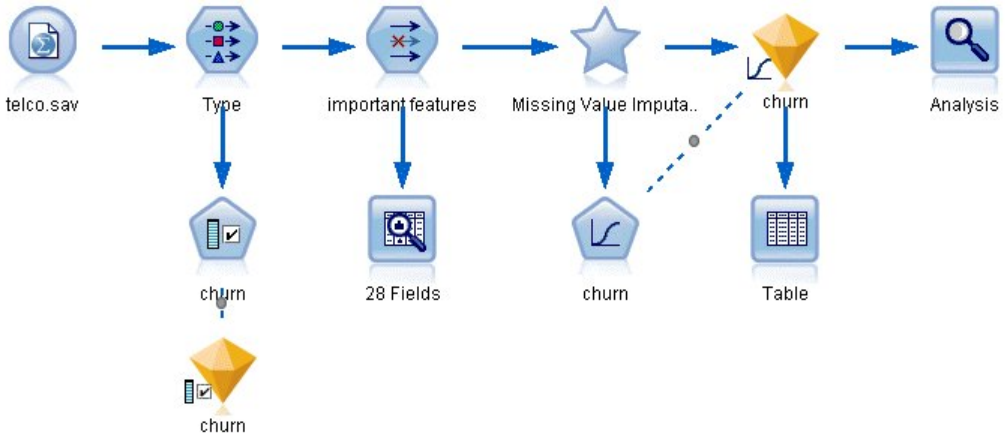


Figura 154. Fluxo de amostra para classificar os clientes usando regressão logística binomial

2. Inclua um nó Tipo para definir campos, certificando-se de que todos os níveis de medição estão configurados corretamente. Por exemplo, a maioria dos campos com valores 0 e 1 pode ser considerada como sinalizadores, mas certos campos, como o gênero, são visualizados com mais precisão como um campo nominal com dois valores.

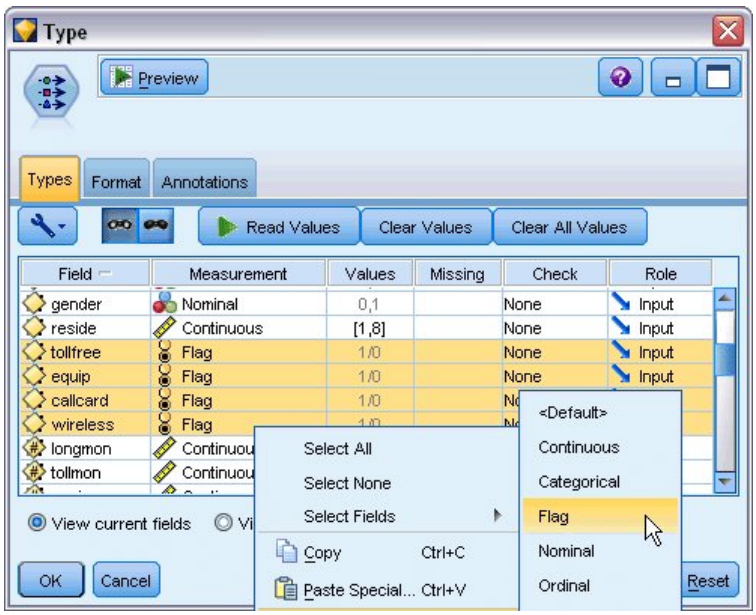


Figura 155. Configurando o nível de medição para vários campos

Dica: Para alterar propriedades para vários campos com valores semelhantes (como 0/ 1), clique no cabeçalho da coluna *Valores* para classificar campos por valor e, em seguida, prenda a tecla Shift enquanto utiliza as teclas do mouse ou seta para selecionar todos os campos que você deseja alterar. Em seguida, é possível clicar com o botão direito do mouse na seleção para alterar o nível de medição ou outros atributos dos campos selecionados.

3. Configure o nível de medição do campo de *rotatividade* como **Sinalizador** e configure a função como **Destino**. Todos os outros campos devem ter seu papel configurado como **Entrada**.

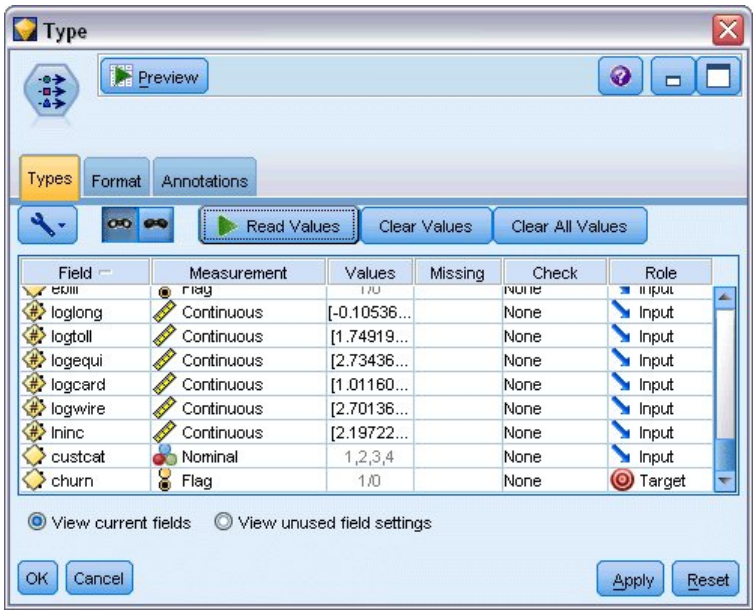


Figura 156. Configurando o nível de medição e o papel para o campo de churn

4. Inclua um nó de modelagem Seleção de recurso no nó Tipo.

   O uso de um nó de Seleção de Recurso possibilita remover preditores ou dados que não adicionem nenhuma informação útil com relação ao relacionamento previda/alvo.

5. Execute o fluxo.
6. Abra o nugget de modelo resultante, e a partir do menu **Gerar** , escolha **Filtrar** para criar um nó Filtro.

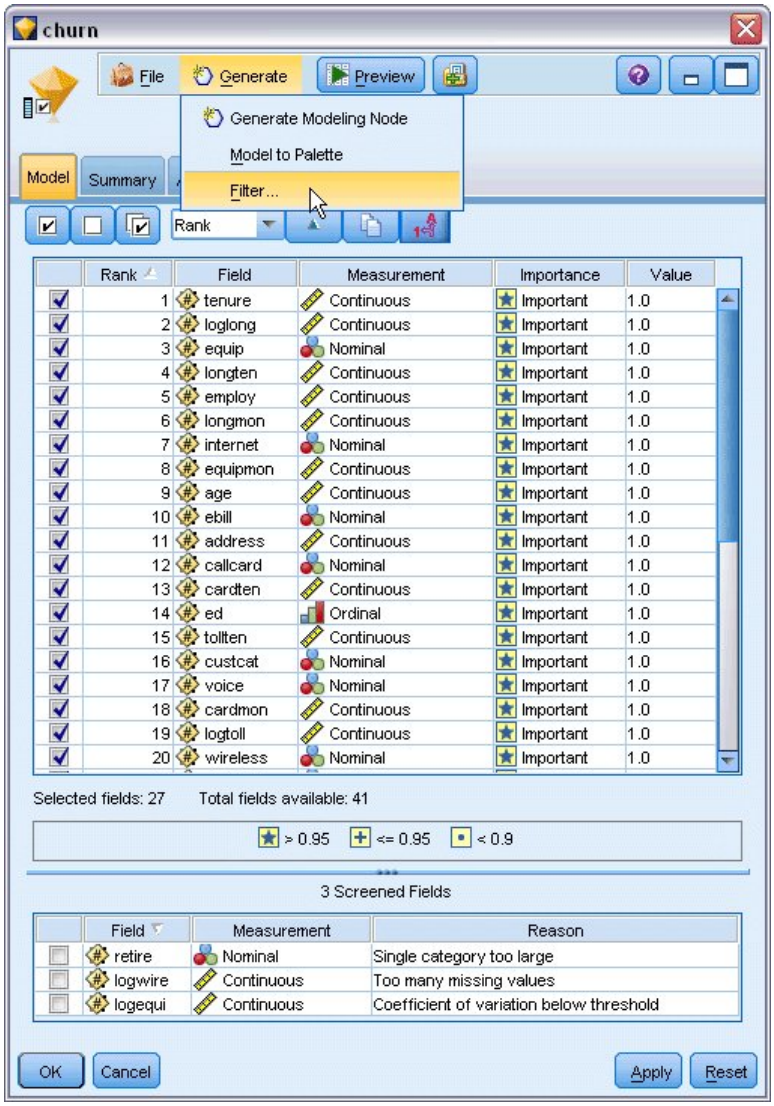


Figura 157. Gerando um Nó Filtro a partir de um nó do Feature Selection

Nem todos os dados no arquivo *telco.sav* serão úteis na previsão de perda de clientes. Você pode usar o filtro para apenas selecionar dados considerados importantes para uso como um preditor.

7. Na caixa de diálogo Gerar Filtro, selecione **Todos os campos marcados: Importante** e clique em **OK**.
8. Conecte o nó do Filtro gerado ao nó Type.

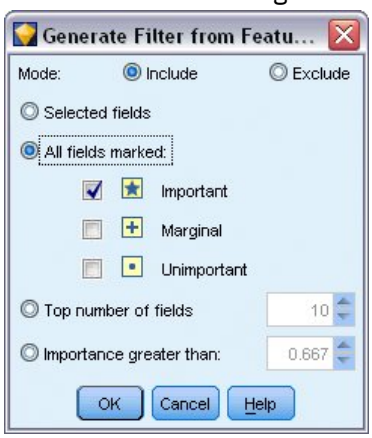


Figura 158. Seleção de campos importantes

9. Anexar um nó de Auditoria De Dados ao nó do Filtro gerado.

Abra o nó de auditoria de Dados e clique em **Executar**.

10. Na guia Qualidade do navegador Data Audit, clique na coluna *% Completo* para classificar a coluna por ordem numérica crescente. Isso permite identificar quaisquer campos com grandes quantidades de dados ausentes; neste caso o único campo que você precisa alterar é *logtoll*, que é menos de 50% completo.
11. Na coluna *Impute Ausente* para *logtoll*, clique em **especificar**.

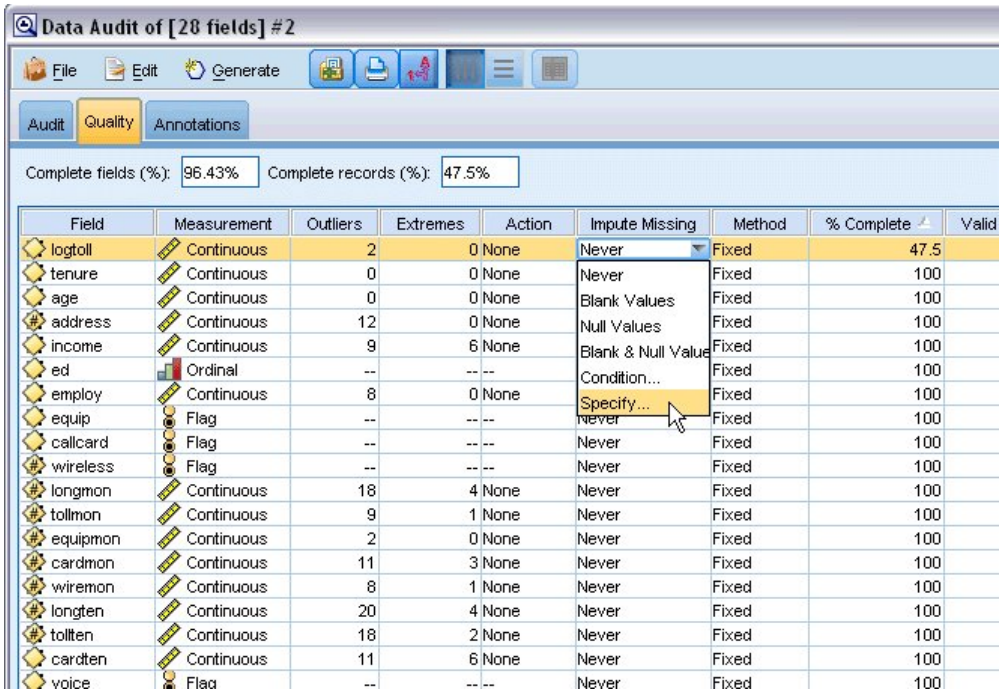


Figura 159. Imputação de valores perdidos para o logado

12. Para **Impute quando**, selecione **Valores Blank e Null**. Para **Fixo Como**, selecione **Mean** e clique em **OK**.

    A seleção de **Mean** garante que os valores imputados não afetem negativamente a média de todos os valores nos dados gerais.

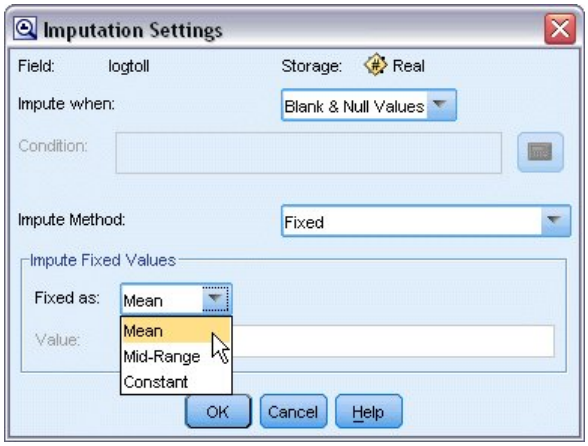


Figura 160. Seleção de configurações de imputação

13. Na guia Qualidade do navegador de Auditoria de Dados, gere os Valores Omistos SuperNode Para fazer isso, a partir dos menus escolha:

    Gerar > **Valores Omissos SuperNode**

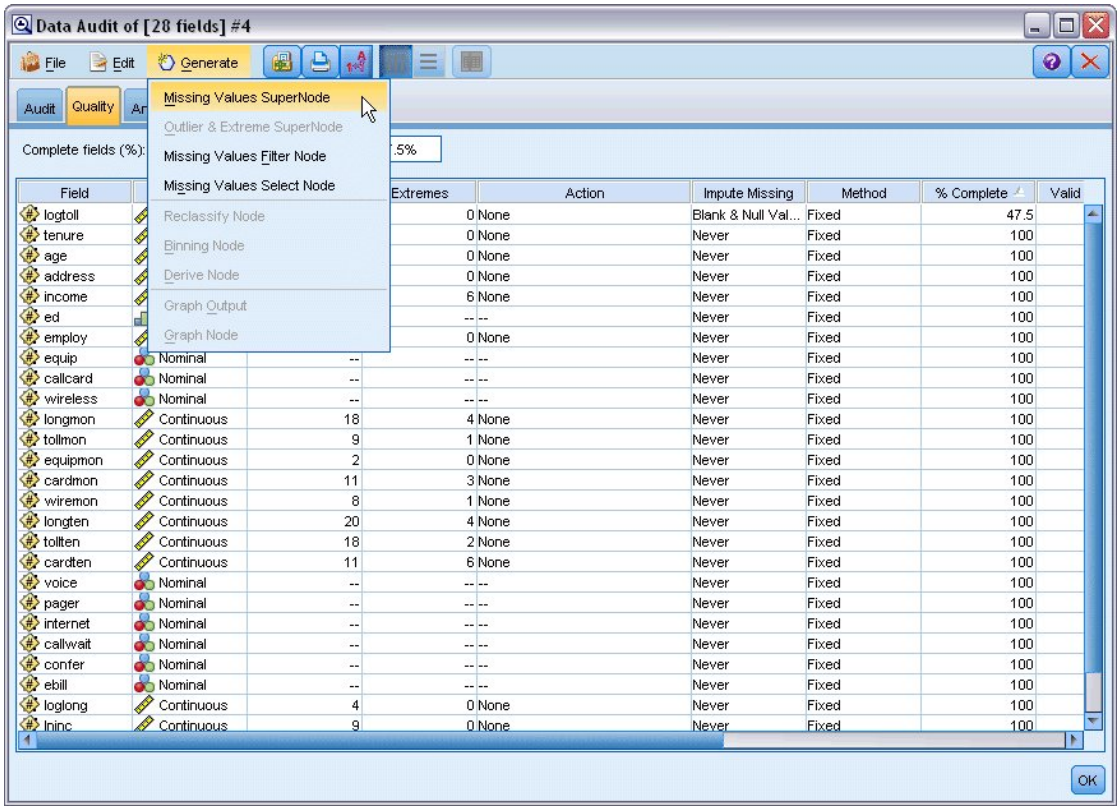


Figura 161. Gerando um valor omisso SuperNode

Na caixa de diálogo Valores Ausentes SuperNode , aumente o **Tamanho de Amostra** para 50% e clique em **OK**.

O SuperNode é exibido na tela de fluxo, com o título: *imputação de valor omisso*.

14. Conecte o SuperNode ao nó de Filtro

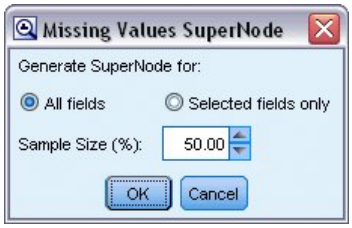


Figura 162. Especificando o tamanho da amostra

15. Inclua um nó Logístico para o SuperNode
16. No nó Logístico, clique na guia Modelo e selecione o procedimento **Binomial** . Na área *Procedimento Binomial* , selecione o método **Forwards** .

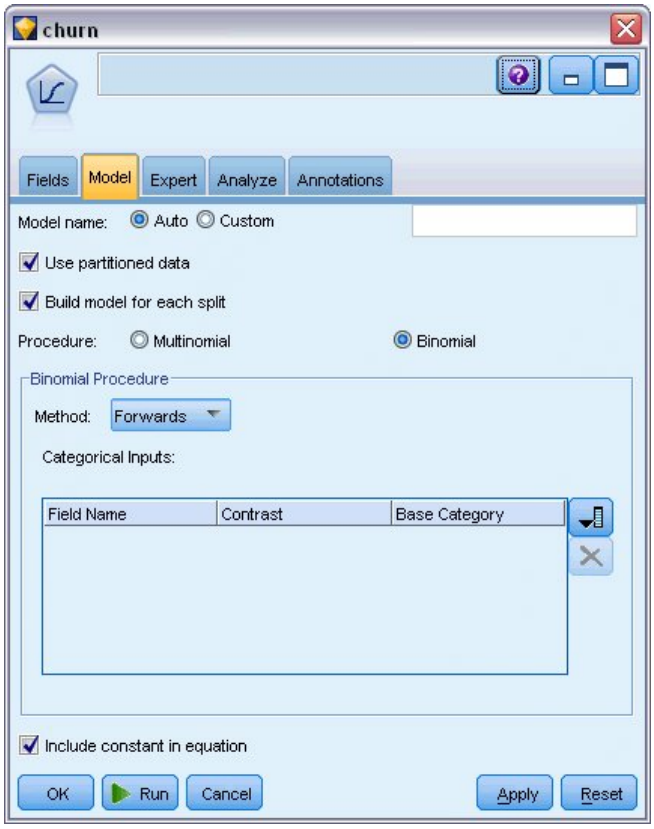


Figura 163. Escolhendo opções de modelo

17. Na guia Expert, selecione o modo **Expert** e, em seguida, clique em **Output**. A caixa de diálogo de Saída Avançada é exibida.
18. No diálogo de Saída Avançada, selecione **Em cada etapa** como o tipo *Display* . Selecione **Histórico de Iteração** e **estimativas de Parâmetro** e clique em **OK**.

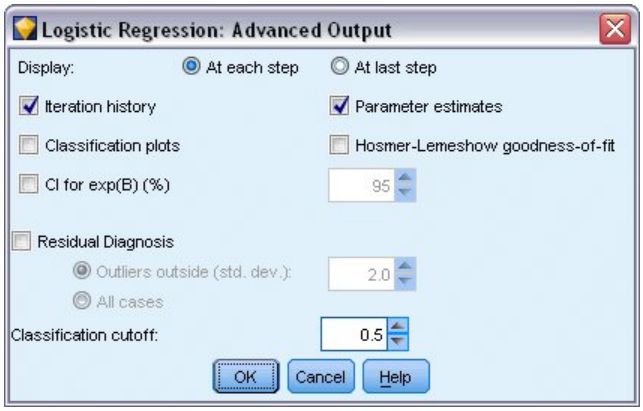


Figura 164. Escolhendo opções de saída

Procurando o Modelo

1. No nó Logístico, clique em **Executar** para criar o modelo.

   O nugget modelo é adicionado à tela do fluxo, e também à paleta de Models no canto superior direito. Para visualizar seus detalhes, clique com o botão direito do mouse sobre o nugget do modelo e selecione **Editar** ou **Browse**.

A guia Resumo mostra (entre outras coisas) o destino e entradas (campos do preditor) utilizados pelo modelo. Observe que estes são os campos que foram realmente escolhidos com base no método Forwards, não a lista completa enviada para consideração.

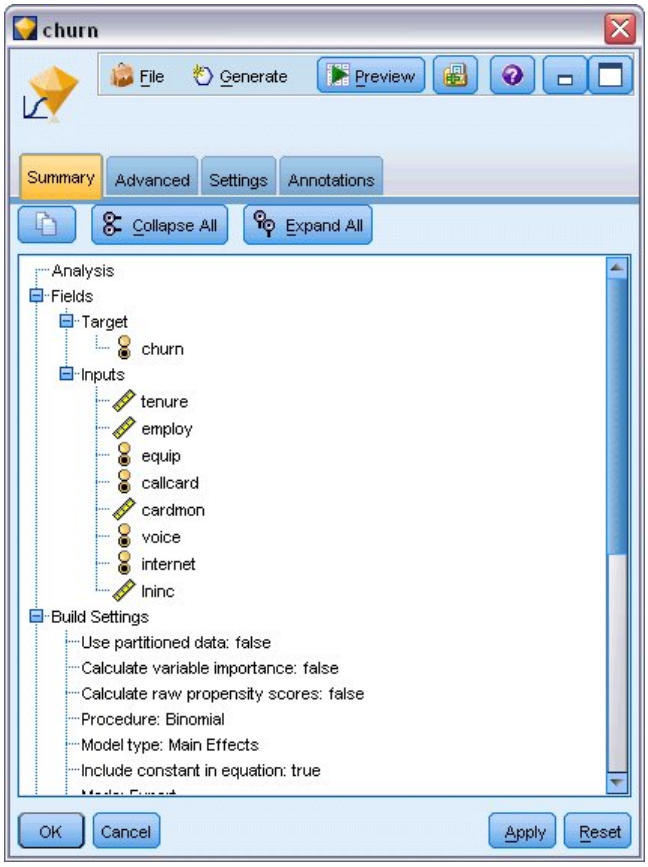


Figura 165. Sumário modelo mostrando campos de destino e entrada

Os itens mostrados na guia Avançado dependem das opções selecionadas na caixa de diálogo de Saída Avançada no nó Logístico. Um item que sempre é mostrado é o Sumário de Processamento de Caso, que mostra o número e a porcentagem de registros incluídos na análise. Além disso, ele lista o número de casos ausentes (se houver) em que um ou mais dos campos de entrada estão indisponíveis e quaisquer casos que não foram selecionados.

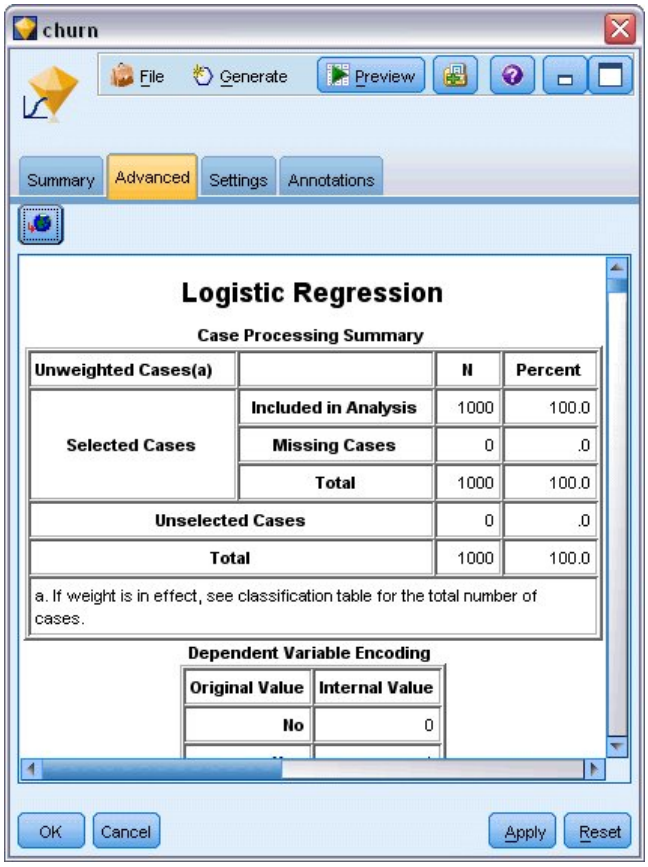


Figura 166. Sumarização do processamento de caso

2. Role a descer do Sumário de Processamento de Caso para exibir a Tabela de Classificação sob o Bloco 0: Bloco Iniciante.

   O Forward Stepwise método começa com um modelo nulo-ou seja, um modelo sem nenhum preditor-que pode ser usado como base para comparação com o modelo construído final. O modelo nulo, por convenção, prevê tudo como um 0, portanto, o modelo nulo é 72.6% preciso simplesmente porque os 726 clientes que não migraram são preditos corretamente. No entanto, os clientes que fizeram churn não estão previstos corretamente em tudo.

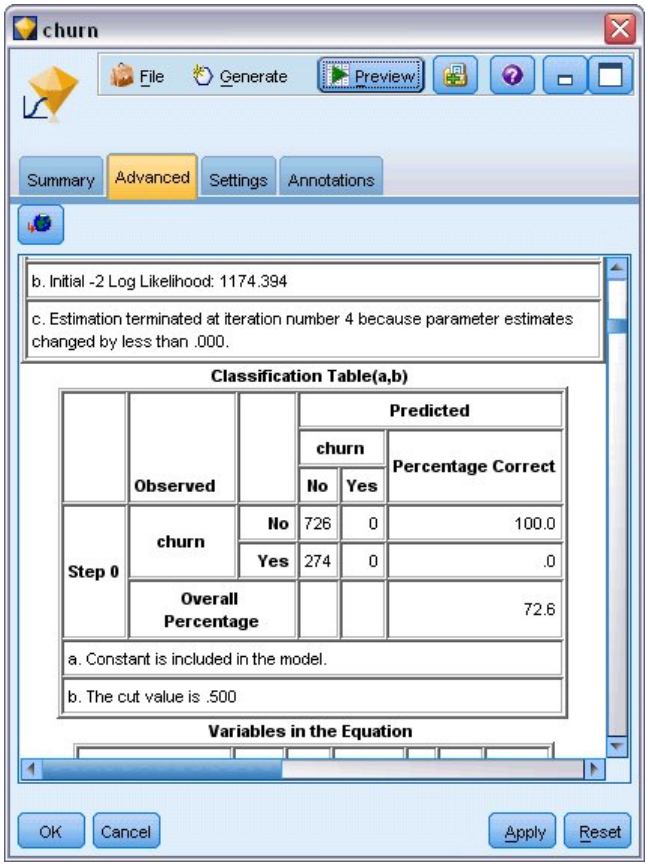


Figura 167. Tabela de classificação inicial-Bloco 0

3. Agora role para baixo para exibir a Tabela de Classificação sob o Bloco 1: Método = Avante Stepwise.

   Esta Tabela de Classificação mostra os resultados para o seu modelo como um preditor é adicionado em cada uma das etapas. Já na primeira etapa-após apenas um preditor ter sido usado-o modelo aumentou a precisão da predição de perda de clientes de 0.0% para 29.9%

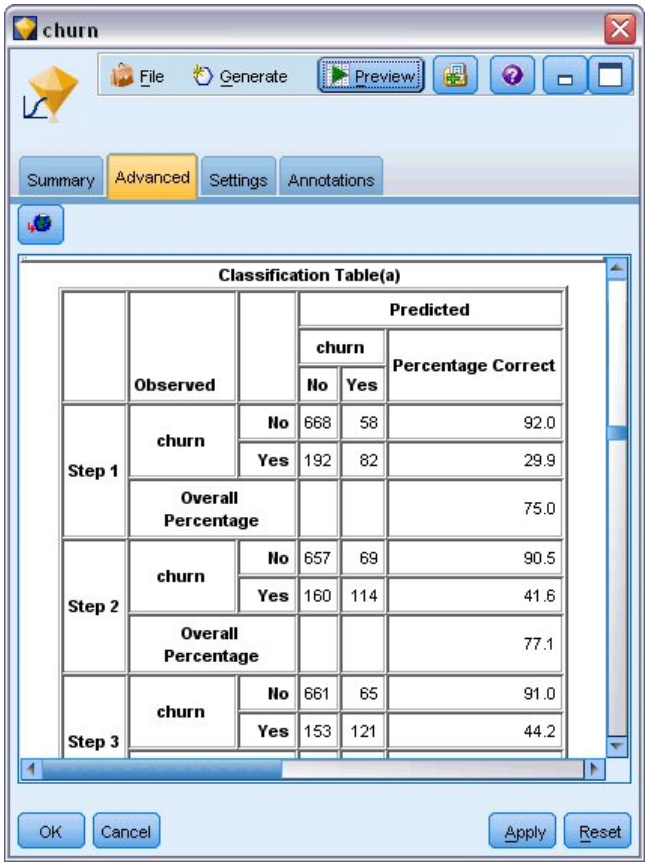


Figura 168. Tabela de classificação-Bloco 1

4. Role para baixo para a parte inferior desta Tabela de Classificação.

A Tabela Classificação mostra que o último passo é o passo 8. Nesta fase o algoritmo decidiu que ele não precisa mais adicionar mais nenhum preditores no modelo. Embora a precisão dos clientes sem perda de clientes tenha diminuído um pouco para 91.2%, a precisão da previsão para aqueles que tiveram perda de clientes aumentou de 0% original para 47.1%. Trata-se de uma melhoria significativa sobre o modelo nulo original que não usava preditores.

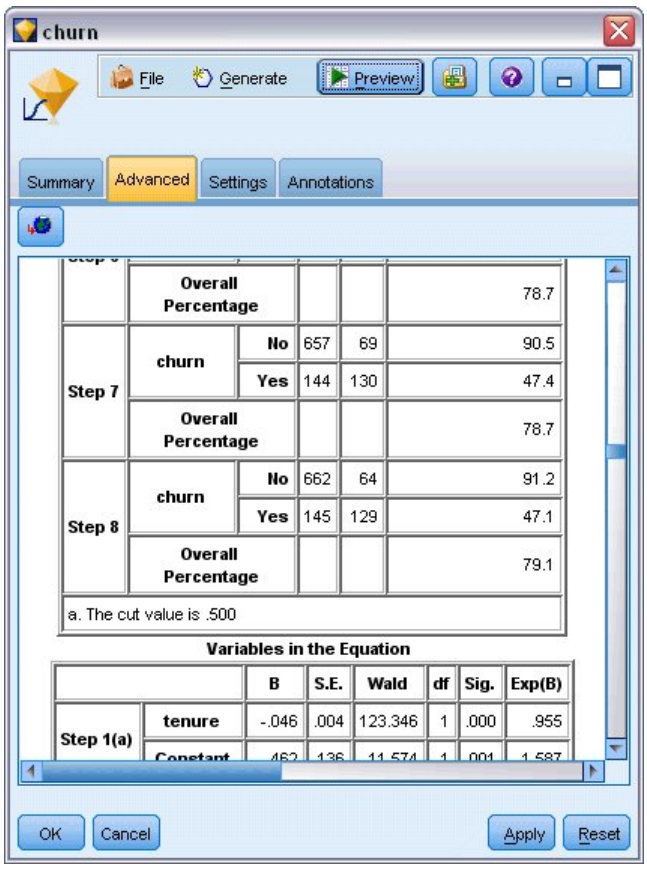


Figura 169. Tabela de classificação-Bloco 1

Para um cliente que deseja reduzir o churn, ser capaz de reduzi-lo por quase metade seria um grande passo na proteção de seus fluxos de renda.

Nota: Este exemplo também mostra como levar o Percentual Geral como um guia para a precisão de um modelo pode, em alguns casos, ser enganoso. O modelo nulo original era 72.6% de precisão geral, enquanto o modelo previsto final tem uma precisão geral de 79.1%; no entanto, como vimos, a precisão das predições de categoria individuais reais eram muito diferentes.

Para avaliar o quão bem o modelo realmente se encaixa nos dados, vários diagnósticos estão disponíveis na caixa de diálogo Advanced Output quando você está construindo o modelo. Explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no *IBM SPSS Modelador Algorithms Guide*, disponível a partir do diretório *Documentação* do disco de instalação.

Observe também que esses resultados são baseados apenas nos dados de treinamento. Para avaliar o quão bem o modelo generaliza para outros dados no mundo real, você usaria um nó de partição para conter um subconjunto de registros para fins de teste e validação.

Capítulo 14. Previsão De Utilização Da Largura De Banda (Série A Tempo)

Previsão com o nó Série Temporal

Um analista de um provedor de banda larga nacional é necessário para produzir previsões e assinaturas do usuário para prever a utilização da largura de banda. As previsões são necessárias para cada um dos mercados locais que formam a base do assinante nacional. Você usará a modelagem de séries temporais para produzir previsões para os próximos três meses para vários mercados locais. Um segundo exemplo mostra como é possível converter dados de origem se ele não estiver no formato correto para entrada para o nó da Série Tempo.

Estes exemplos usam o fluxo denominado *broadband_create_models.str*, que faz referência ao arquivo de dados denominado *broadband_1.sav* Esses arquivos estão disponíveis a partir da pasta *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *broadband_create_models.str* está na pasta *fluxos* .

O último exemplo demonstra como aplicar os modelos salvos em um dataset atualizado a fim de estender as previsões por mais três meses.

Em IBM SPSS Modelador, é possível produzir diversos modelos de séries temporais em uma única operação. O arquivo de origem que você estará usando tem dados de séries temporais para 85 mercados diferentes, embora para o bem da simplicidade você só irá modelar cinco desses mercados, mais o total para todos os mercados.

O arquivo de dados *broadband_1.sav* tem dados de uso mensais para cada um dos 85 mercados locais. Para os fins deste exemplo, apenas as cinco primeiras séries serão usadas; um modelo separado será criado para cada uma dessas cinco séries, mais um total.

O arquivo também inclui um campo de data que indica o mês e o ano de cada registro. Este campo será usado para rotular registros. O campo date lê em IBM SPSS Modelador como uma string, mas, para usar o campo em IBM SPSS Modelador você irá converter o tipo de armazenamento em formato numérico Data usando um nó Filler.

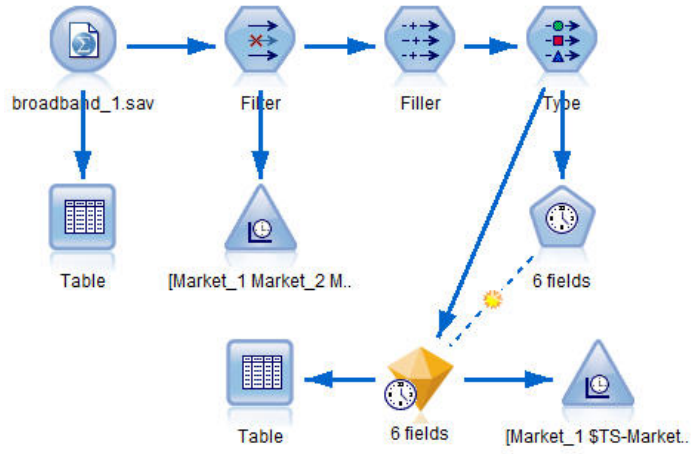
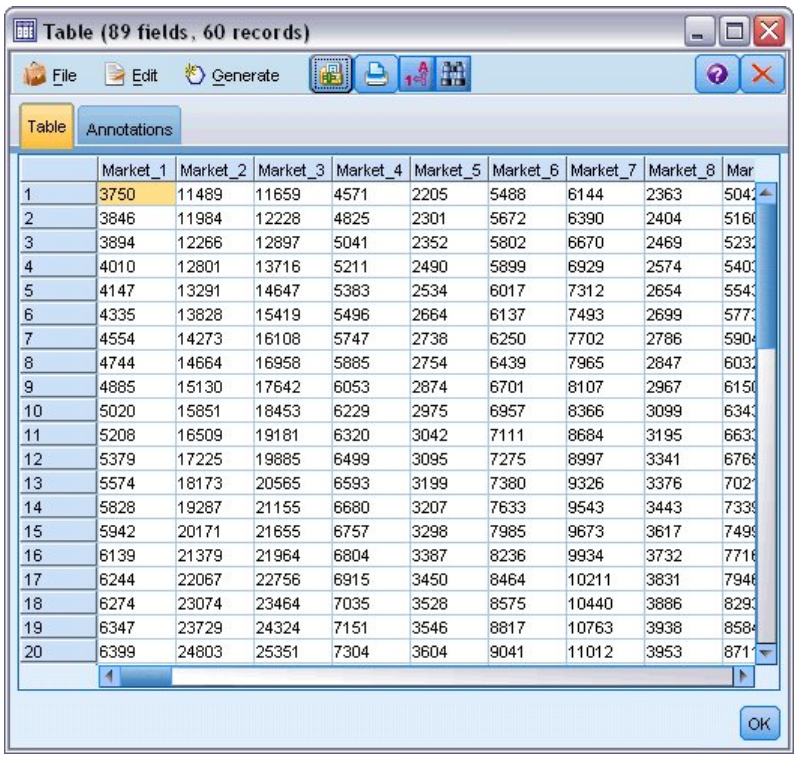


Figura 170. Fluxo de amostra para mostrar modelagem da Série Time

O nó da Série Tempo requer que cada série esteja em uma coluna separada, com uma linha para cada intervalo. IBM SPSS Modelador fornece métodos para transformar dados para combinar com este formato se necessário.



	Market_1	Market_2	Market_3	Market_4	Market_5	Market_6	Market_7	Market_8	Mar
1	3750	11489	11659	4571	2205	5488	6144	2363	5042
2	3846	11984	12228	4825	2301	5672	6390	2404	5160
3	3894	12266	12897	5041	2352	5802	6670	2469	5232
4	4010	12801	13716	5211	2490	5899	6929	2574	5403
5	4147	13291	14647	5383	2534	6017	7312	2654	5543
6	4335	13828	15419	5496	2664	6137	7493	2699	5773
7	4554	14273	16108	5747	2738	6250	7702	2786	5904
8	4744	14664	16958	5885	2754	6439	7965	2847	6032
9	4885	15130	17642	6053	2874	6701	8107	2967	6150
10	5020	15851	18453	6229	2975	6957	8366	3099	6343
11	5208	16509	19181	6320	3042	7111	8684	3195	6633
12	5379	17225	19885	6499	3095	7275	8997	3341	6769
13	5574	18173	20565	6593	3199	7380	9326	3376	7021
14	5828	19287	21155	6680	3207	7633	9543	3443	7339
15	5942	20171	21655	6757	3298	7985	9673	3617	7499
16	6139	21379	21964	6804	3387	8236	9934	3732	7716
17	6244	22067	22756	6915	3450	8464	10211	3831	7948
18	6274	23074	23464	7035	3528	8575	10440	3886	8293
19	6347	23729	24324	7151	3546	8817	10763	3938	8584
20	6399	24803	25351	7304	3604	9041	11012	3953	8711

Figura 171. Dados de assinatura mensal para mercados locais de banda larga

Criando o Fluxo

1. Crie um novo fluxo e inclua um nó de origem do File File apontando para *broadband_1.sav*.
2. Use um nó Filtro para filtrar os campos *Market_6* para *Market_85* e os campos *MONTH_* e *ANO_* para simplificar o modelo.

Dica: Para selecionar vários campos adjacentes em uma única operação, clique no campo *Market_6* , aguça o botão esquerdo do mouse e arraste o mouse para baixo no campo *Market_85* . Os campos selecionados são destacados em azul. Para adicionar os outros campos, mantenha pressionada a tecla Ctrl e clique nos campos *MONTH_* e *YEAR_* .

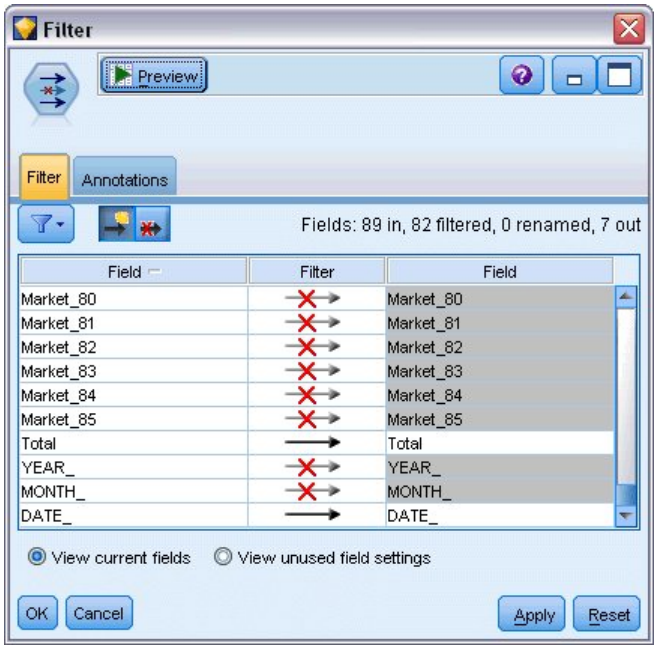


Figura 172. Simplificando o modelo

Examinando os dados

É sempre uma boa ideia ter uma sensação para a natureza dos seus dados antes de construir um modelo. Os dados apresentam variações sazonais? Embora o Expert Modeler possa encontrar automaticamente o melhor modelo sazonal ou não sazonal para cada série, você pode, muitas vezes, obter resultados mais rápidos limitando a busca a modelos não sazonais quando a sazonalidade não está presente em seus dados. Sem examinar os dados de cada um dos mercados locais, podemos obter uma imagem aproximada da presença ou ausência de sazonalidade ao plotar o número total de assinantes em todos os cinco mercados.

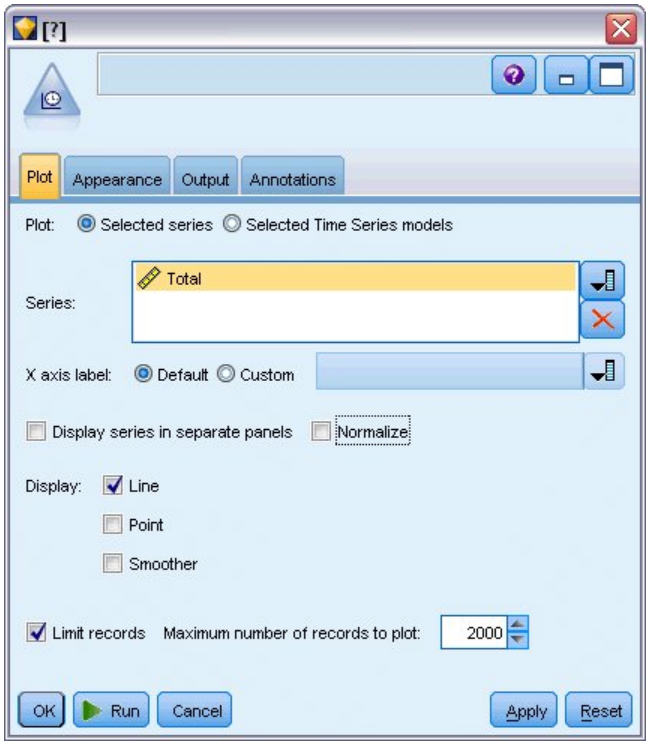


Figura 173. Plotar o número total de assinantes

1. Na paleta Gráficos, anexe um nó do gráfico de tempo ao nó de filtro.
2. Inclua o campo *Total* na lista Série.
3. Desmarque a **Série Exibir em painéis separados** e **Normalize** caixas de seleção.
4. Clique em **Executar** .

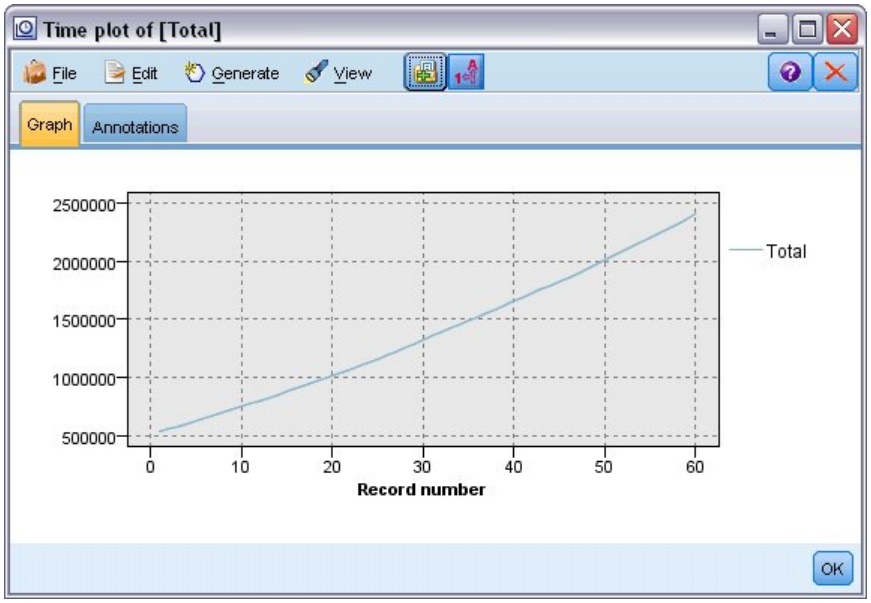


Figura 174. Trama temporal do campo Total

A série exibe uma tendência crescente muito suave, sem indícios de variações sazonais. Pode haver séries individuais com sazonalidade, mas parece que a sazonalidade não é uma característica proeminente dos dados em geral.

É claro que você deve inspecionar cada uma das séries antes de descartar modelos sazonais. É possível então separar as séries que exibem sazonalidade e modelá-las separadamente.

O IBM SPSS Modelador torna mais fácil plotar várias séries juntas.

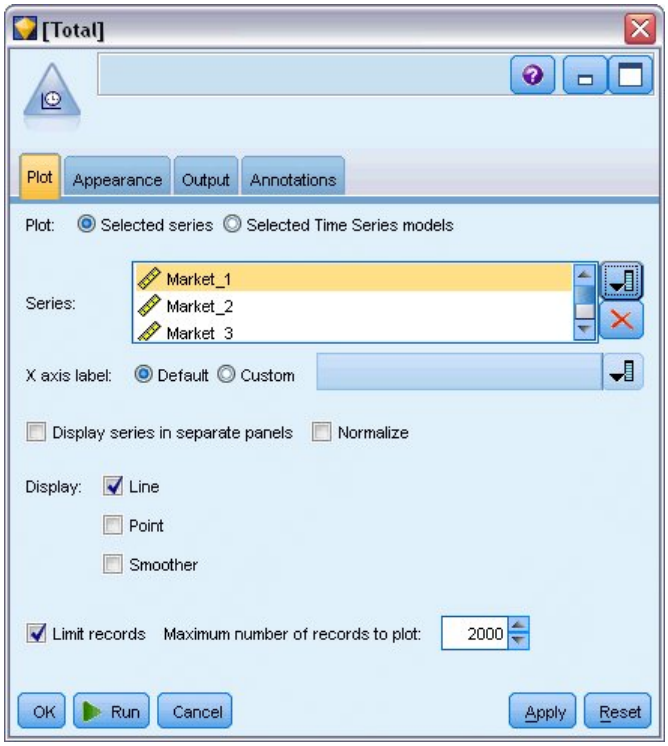


Figura 175. Plotar várias séries temporais

5. Reabra o nó do Time Plot.
6. Remova o campo *Total* da lista da Série (selecione-o e clique no botão vermelho X).
7. Inclua o *Market_1* por meio dos campos *Market_5* na lista.
8. Clique em **Executar** .

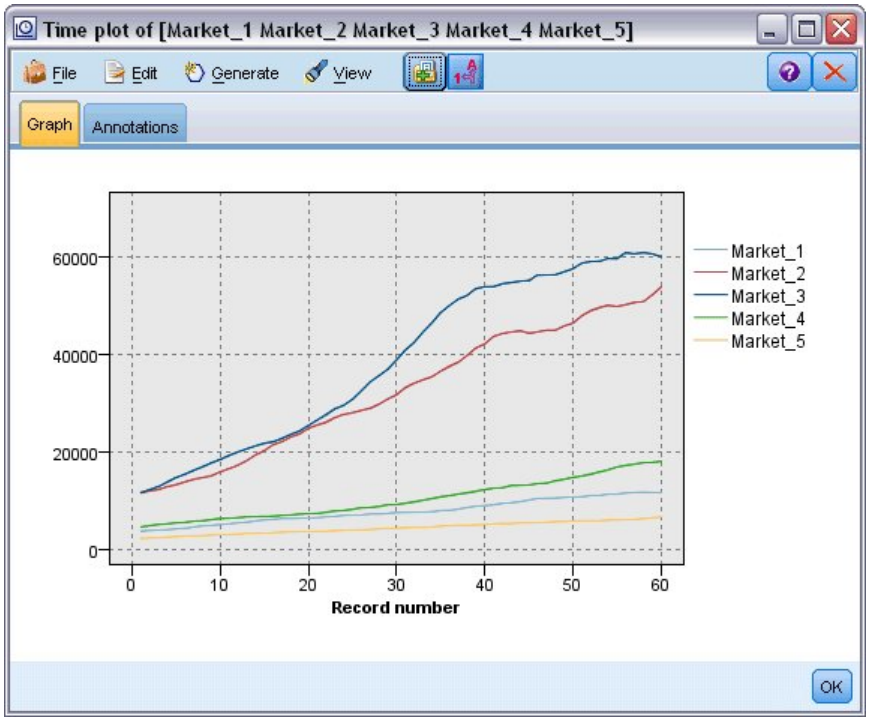


Figura 176. Gráfico de tempo de vários campos

A inspeção de cada um dos mercados revela uma tendência crescente constante em cada caso. Embora alguns mercados sejam um pouco mais erráticos do que outros, não há evidência de sazonalidade a ser vista.

Definindo as datas

Agora é necessário mudar o tipo de armazenamento do campo *DATE_* para o formato de data.

1. Conecte um nó Filler ao nó do Filtro.
2. Abra o nó Filler e clique no botão seletor de campo.
3. Selecione **DATE_** para adicioná-lo em **Fill in fields**.
4. Configure a condição **Substituir** para **Sempre**.
5. Configure o valor de **Substituir com** para **to_date (DATE_)**.

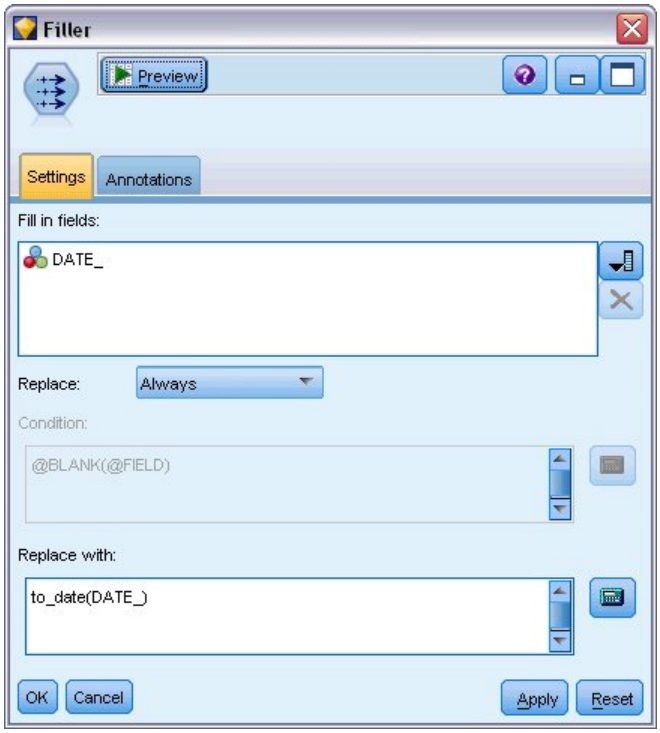


Figura 177. Configurando o tipo de armazenamento de dados

Alterar o formato de data padrão para combinar com o formato do campo Data. Isso é necessário para que a conversão do campo Data funcione conforme o esperado.

6. No menu, escolha **Ferramentas> Propriedades do Stream> Opções** para exibir a caixa de diálogo Opções de Fluxo.
7. Selecione a pane **Data / Hora** e configure o padrão de **Data format** para **MON YYYY** .

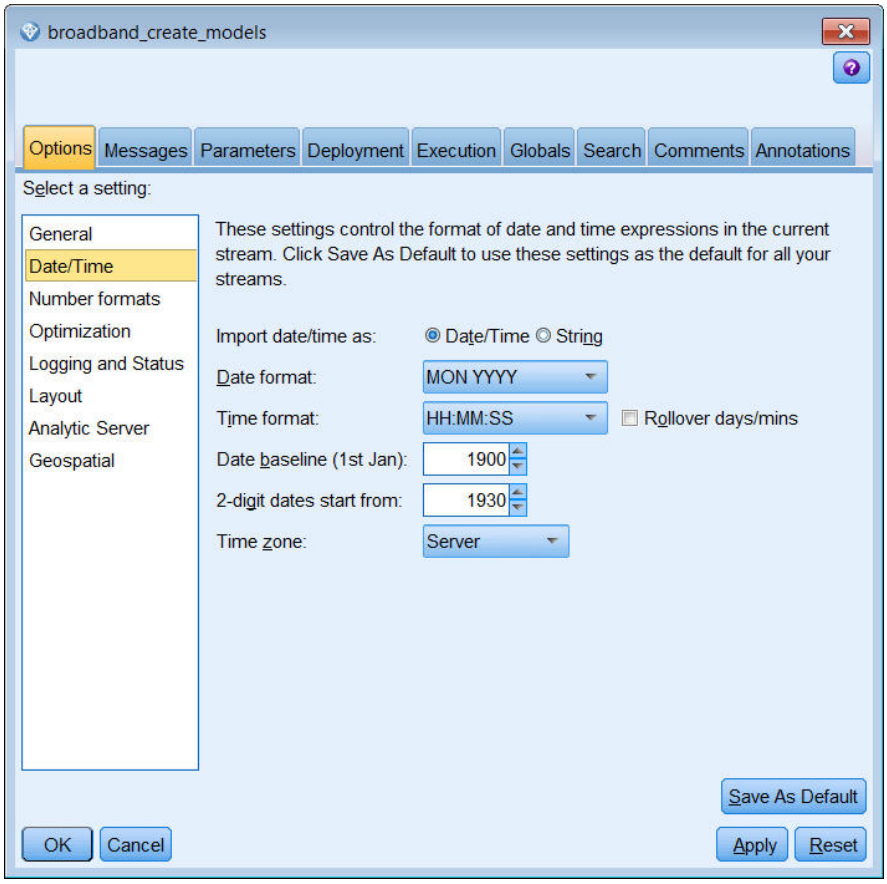


Figura 178. Configurando o formato de data

Definindo os destinos

1. Inclua um nó Type e configure a função para **Nenhum** para o campo *DATE_* . Configure a função para **Target** para todos os outros (os campos *Market_n* mais o campo *Total*).
2. Clique no botão **Ler Valores** para preencher a coluna Valores.

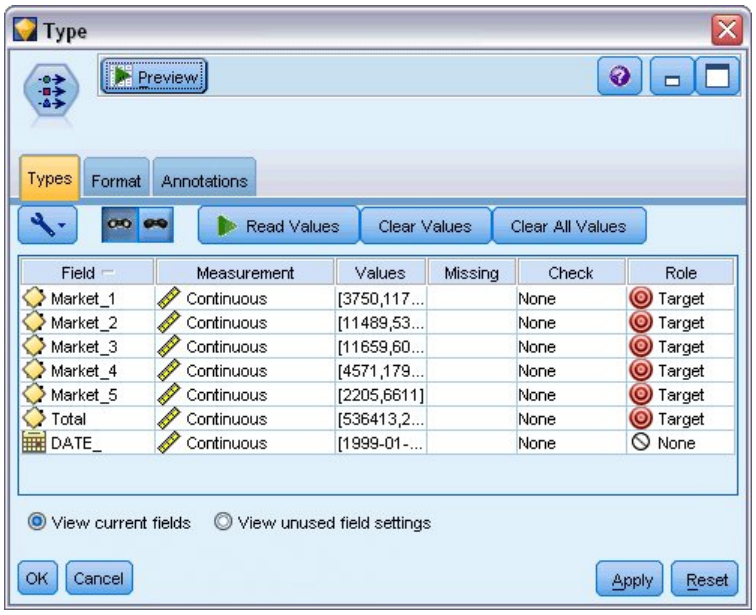


Figura 179. Configurando a função para vários campos

Configurando os intervalos de tempo

1. A partir da paleta Modelagem, adiciona um nó do Time Series ao stream e anexe-o ao nó Type.
2. Na guia Especificações de Dados, na área de janela de Observações, selecione DATE_ como o campo **Data / hora**.
3. Selecione Months como o **Intervalo de tempo**.

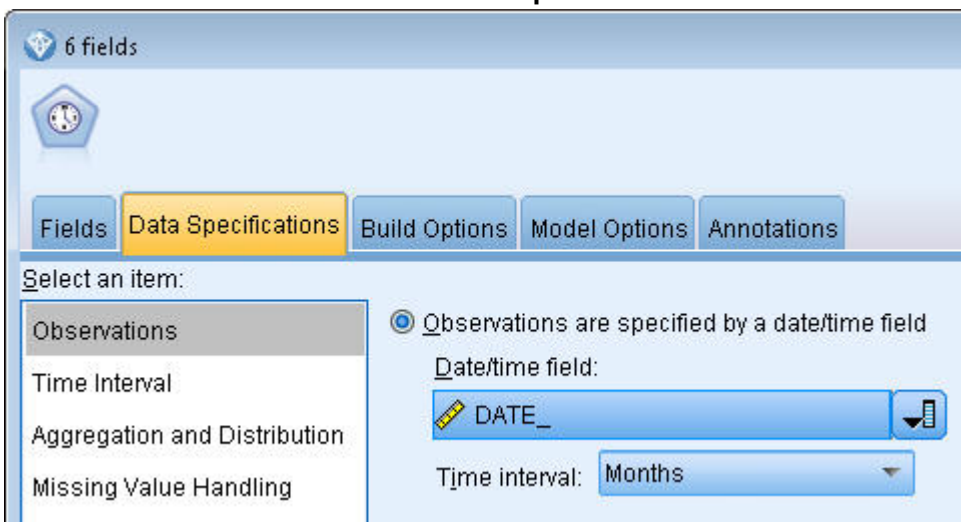


Figura 180. Configurando o intervalo de tempo

4. Na guia Opções do Modelo, selecione os **Registros de Extensão no futuro** caixa de seleção.
5. Configure o valor para **3**.

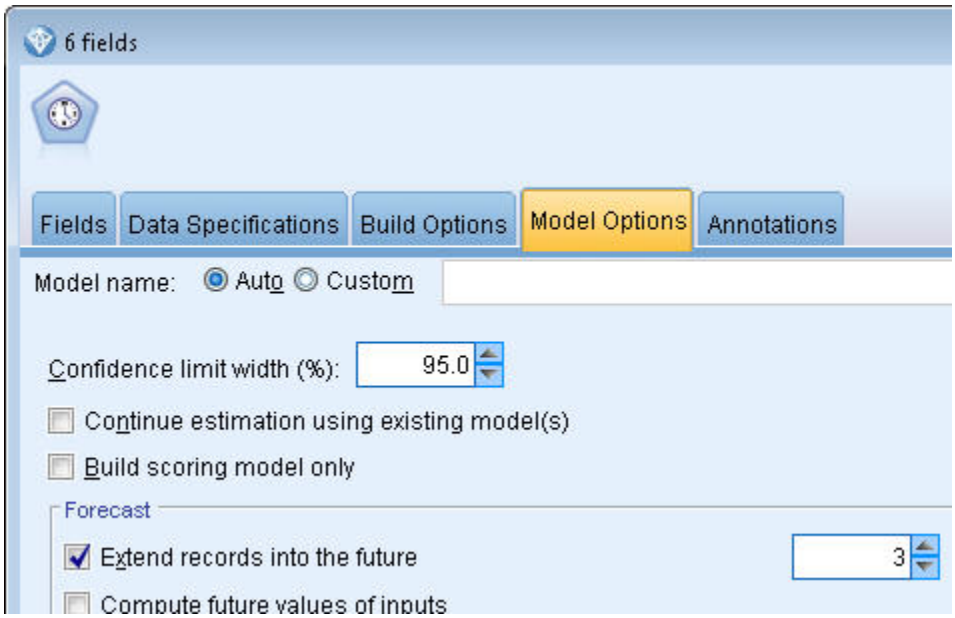


Figura 181. Configurando o período de previsão

Criando o modelo

1. No nó da Série Tempo, escolha a guia Campos. Na lista **Campos** , selecione todos os 5 dos mercados e copie-os para as listas **Destinos de Targets** e **Candidatas** . Além disso, selecione e copie o campo Total para a lista **Targets** .
2. Escolha a guia Opções de Construção e, na pane Geral, certifica-se o Método Expert Modeler **Método** é selecionado usando todas as configurações padrão. Isso permite que o Expert Modeler decida o modelo mais apropriado a ser usado para cada série temporal. Clique em **Executar**.

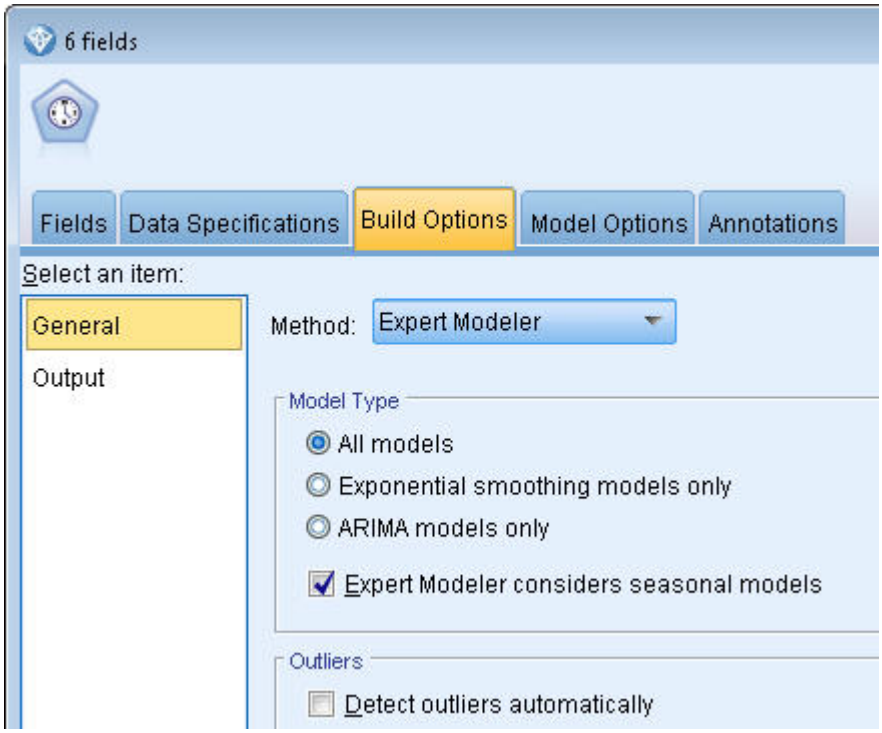


Figura 182. Como escolher o Expert Modeler for Time Series

3. Conecte o nugget do modelo Time Series ao nó da Série Time.
4. Conecte um nó da Tabela ao nugget do modelo Time Series e clique em **Executar**.

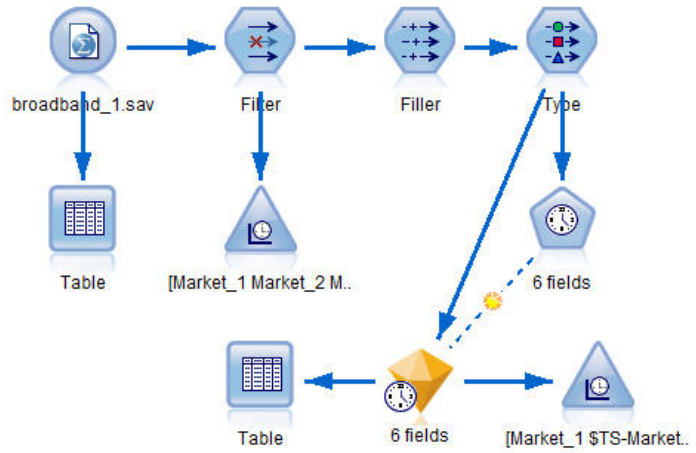


Figura 183. Fluxo de amostra para mostrar modelagem da Série Time

Existem agora três novas linhas (61 até 63) anexadas aos dados originais. Estas são as linhas para o período de previsão, neste caso janeiro a março de 2004.

Várias colunas novas também estão presentes agora; as colunas *$TS-* são adicionadas pelo nó do Time Series. As colunas indicam o seguinte para cada linha (ou seja, para cada intervalo nos dados da série temporal):

Coluna	**Descrição**
$TS-*colname*	Os dados do modelo gerados para cada coluna dos dados originais.
$TSLCI-*colname*	O valor de intervalo de confiança inferior para cada coluna dos dados do modelo gerado.

Coluna	Descrição
$TSUCI-*colname*	O valor de intervalo de confiança superior para cada coluna dos dados do modelo gerado.
$TS-Total	O total dos valores $TS-*colname* para essa linha.
$TSLCI-Total	O total dos valores $TSLCI-*colname* para esta linha.
$TSUCI-Total	O total dos valores $TSUCI-*colnome* para esta linha.

As colunas mais significativas para a operação de previsão são as colunas *$TS-Market_n*, *$TSLCI-Market_n* e *$TSUCI-Market_n*. Em particular, essas colunas nas linhas de 61 63 contêm os dados de previsão de assinatura do usuário e intervalos de confiança para cada um dos mercados locais.

Examinando o modelo

1. Clique duas vezes no nugget do modelo Time Series, e selecione a guia Output para exibir dados sobre os modelos gerados para cada um dos mercados.

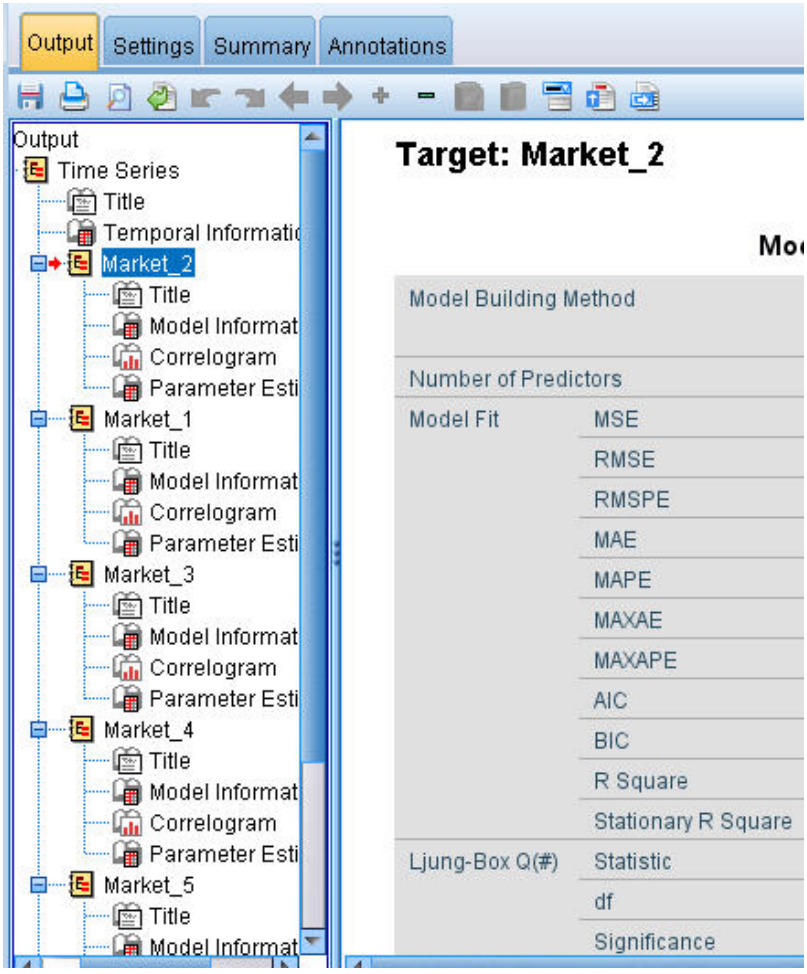


Figura 184. Modelos de séries temporais gerados para os mercados

Na coluna Saída de esquerda, selecione a **Informações do Modelo** para qualquer um dos Mercados. A linha **Número de Preditadores** mostra quantos campos foram usados como preditores para cada destino; neste caso, nenhum.

As linhas restantes nas tabelas **Modelo de Informações** mostram várias medidas de bondade de ajuste para cada modelo. O valor **Estacionário R Square** fornece uma estimativa da proporção da

variação total da série que é explicada pelo modelo. Quanto maior o valor (para um máximo de 1.0), melhor o ajuste do modelo.

As linhas **Q (#) Estatística**, **df**e **Significância** se relacionam com a estatística Ljung-Box, um teste de aleatoriedade dos erros residuais no modelo; quanto mais aleatórios os erros, melhor o modelo provavelmente será. **Q (#)** é a própria estatística Ljung-Box, enquanto **df** (graus de liberdade) indica o número de parâmetros do modelo que são livres para variar ao estimar um determinado destino.

A linha **Significância** dá o valor de significância da estatística Ljung-Box, fornecendo outra indicação de se o modelo está corretamente especificado. Um valor de significância menor que 0.05 indica que os erros residuais não são aleatórios, implicando que há estrutura na série observada que não é contabilizada pelo modelo.

Levando em consideração tanto os valores **Estacionários R Square** quanto **Significância** , os modelos que o Expert Modeler escolheu para *Market_3*, e *Market_4* são bastante aceitáveis. Os valores de **significância** parar *Market_1*, *Market_2* e *Market_5* são todos menores que 0.05, indicando que pode ser necessária alguma experimentação com modelos de melhor ajuste para esses mercados.

O monitor mostra uma série de medidas adicionais de adequação. O valor do **R Square** dá uma estimativa da variação total da série temporal que pode ser explicada pelo modelo. Como o valor máximo para essa estatística é 1.0, nossos modelos são bons nesse aspecto.

RMSE é a raiz do erro quadrático médio, uma medida de quanto os valores reais de uma série diferem dos valores previstos pelo modelo, e é expressa nas mesmas unidades usadas para a própria série. Como esta é uma medição de um erro, queremos que esse valor seja o mais baixo possível. À primeira vista, parece que os modelos para *Market_2* and *Market_3*, embora ainda aceitáveis de acordo com as estatísticas que vimos até agora, têm menos sucesso do que aqueles para os outros três mercados.

Essas medidas adicionais de adequação incluem os erros percentuais absolutos médios (**MAPE**) e seu valor máximo (**MAXAPE**). O erro percentual absoluto é uma medida de quanto uma série de destino varia de seu nível predito pelo modelo, expresso como um valor percentual. Ao examinar a média e o máximo em todos os modelos, é possível obter uma indicação da incerteza em suas previsões.

O valor do MAPE mostra que todos os modelos exibem uma incerteza média de cerca de 1%, o que é muito baixo. O valor MAXAPE exibe o erro de percentual absoluto absoluto e é útil para imaginar um pior cenário para suas previsões. Isso mostra que o maior erro de porcentagem para a maioria dos modelos cai no intervalo de aproximadamente 1.8 a 3.7%, novamente um conjunto muito baixo de números, com apenas *Market_4* sendo mais alto em 7% mais próximo.

O valor de **MAE** (erro médio absoluto) mostra a média dos valores absolutos dos erros de previsão. Assim como o valor de RMSE, isso é expresso nas mesmas unidades que as usadas para a própria série. **MAXAE** mostra o maior erro de previsão nas mesmas unidades e indica o pior cenário para as previsões.

Interessante embora esses valores absolutos sejam, são os valores dos erros de porcentagem (MAPE e MAXAPE) que são mais úteis nesse caso, já que a série alvo representam números de assinantes para mercados de tamanhos variados.

Os valores MAPE e MAXAPE representam uma quantidade aceitável de incerteza com os modelos? Eles são certamente muito baixos. Essa é uma situação em que o bom senso comercial entra em jogo, porque o risco aceitável mudará de problema para problema. Vamos supor que as estatísticas de bondade-de-ajuste caem dentro de limites aceitáveis e vão em frente para observar os erros residuais.

Examinar os valores da função de autocorrelação (ACF) e função de autocorrelação parcial (PACF) para os residuos do modelo fornece visão mais quantitativa sobre os modelos do que simplesmente visualizar estatísticas de bondade-de-ajuste.

Um modelo de série temporal bem especificado capturará toda a variação não aleatória, incluindo sazonalidade, tendência e fatores cíclicos e outros que são importantes. Nesse caso, qualquer erro não deve ser correlacionado consigo mesmo (autocorrelacionado) ao longo do tempo. Uma estrutura

significativa em qualquer uma das funções de autocorrelação implicaria que o modelo subjacente está incompleto.

2. Para o quarto mercado, na coluna da esquerda, clique em **Correlogram** para exibir os valores da função de autocorrelação (ACF) e função de autocorrelação parcial (PACF) para os erros residuais no modelo.

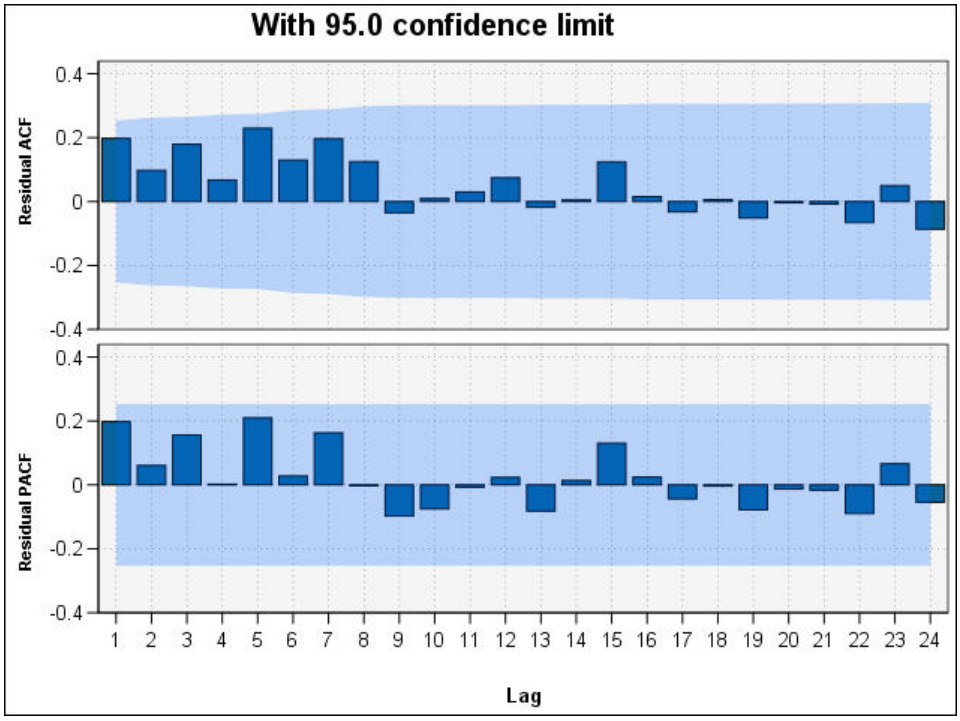


Figura 185. Valores ACF e PACF para o quarto mercado

Nessas tramas, os valores originais da variável de erro foram defasados em até 24 períodos de tempo e comparados com o valor original para ver se há alguma correlação ao longo do tempo. Para que o modelo seja aceitável, nenhuma das barras na trama superior (ACF) deve se estender fora da área sombreada, em uma direção positiva (para cima) ou negativa (para baixo).

Caso isso ocorra, você precisaria verificar a trama inferior (PACF) para ver se a estrutura está confirmada ali. A trama da PACF olha para correlações depois de controlar para os valores da série nos pontos de tempo intervencionista.

Os valores para *Market_4* estão todos dentro da área sombreada, então podemos continuar e verificar os valores para os outros mercados.

3. Clique no **Correlograma** para cada um dos outros mercados e os totais.

   Todos os valores para os outros mercados mostram alguns valores fora da área sombreada, confirmando o que suspeitamos anteriormente de seus valores de **significância**. Teremos de experimentar alguns modelos diferentes para esses mercados em algum ponto para ver se podemos obter um ajuste melhor, mas para o resto deste exemplo, vamos nos concentrar no que mais podemos aprender com o modelo *Market_4*.

4. A partir da paleta de Gráficos, anexe um nó do Time Plot no nugget modelo da Série Time.
5. Na guia Plot, limpe a caixa de seleção **Exibir séries em painéis separados** .
6. Na lista **Series** , clique no botão seletor de campo, selecione os campos *Market_4* e *$TS-Market_4* , e clique em **OK** para incluí-los na lista.
7. Clique em **Executar** para exibir um gráfico de linha dos dados reais e previstos para o primeiro dos mercados locais.

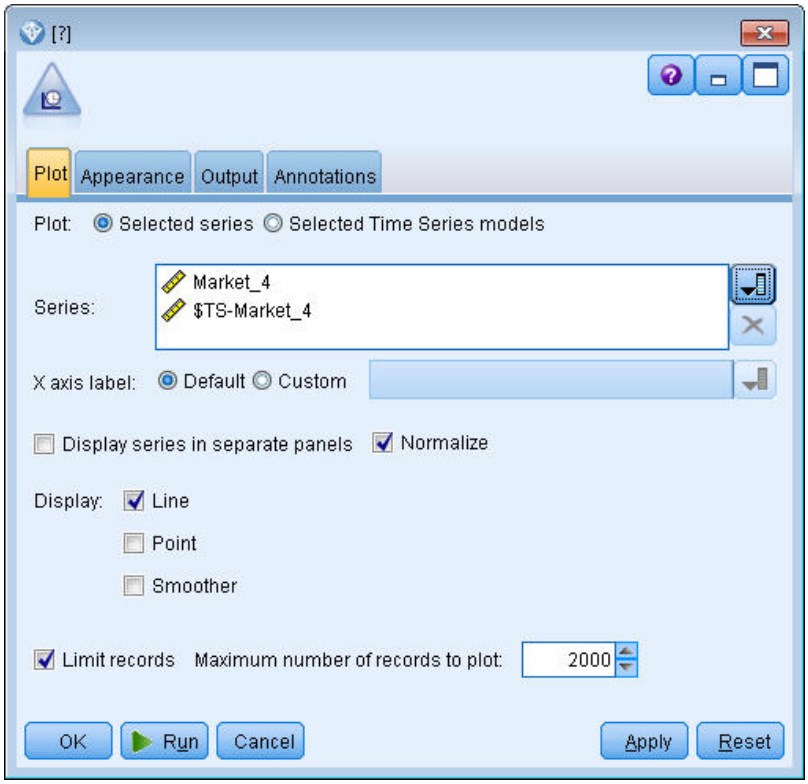


Figura 186. Seleção dos campos para trama

Observe como a linha de previsão (*$TS-Market_4*) se estende além do final dos dados reais. Você agora tem uma previsão de demanda esperada para os próximos três meses neste mercado.

As linhas para dados reais e previstos ao longo de toda a série temporal estão muito próximas no gráfico, indicando que este é um modelo confiável para esta série temporal específica.

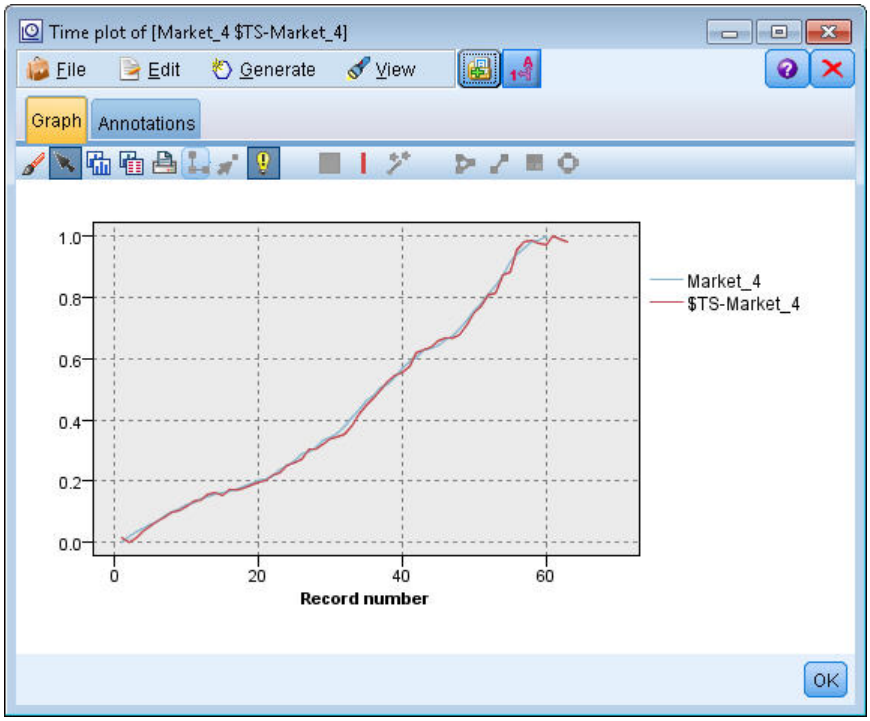


Figura 187. Gráfico de tempo de dados reais e de previsão para Market_4

Salve o modelo em um arquivo para uso em um exemplo futuro:

8. Clique em **OK** para fechar o gráfico atual.
9. Abra o nugget modelo do Time Series.
10. Escolha **Arquivo> Salvar Nó** e especifique o local do arquivo.
11. Clique em **Salvar**.

    Você tem um modelo confiável para este mercado específico, mas qual é a margem de erro que a previsão tem? É possível obter uma indicação disso examinando o intervalo de confiança.
12. Clique duas vezes no último nó do Plot Time no fluxo (aquele rotulado como **Market_4 $TS-Market_4**) para abrir sua caixa de diálogo novamente.
13. Clique no botão seletor de campo e inclua os campos *$TSLCI-Market_4* e *$TSUCI-Market_4* para a lista **Series** .
14. Clique em **Executar** .

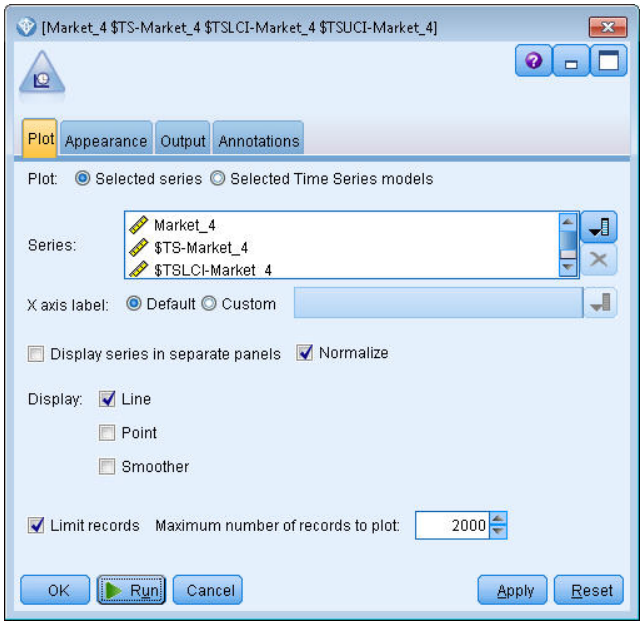


Figura 188. Adicionando mais campos para enredo

Agora você tem o mesmo gráfico de antes, mas com os limites superior (*$TSUCI*) e inferior (*$TSLCI*) do intervalo de confiança incluído.

Observe como os limites do intervalo de confiança divergem ao longo do período de previsão, indicando o aumento da incerteza à medida que você faz previsões no futuro.

No entanto, como cada período de tempo passa, você terá outro valor (neste caso) mês de dados de uso real sobre o qual basear a sua previsão. Você pode ler os novos dados no fluxo e reaplicar o seu modelo agora que você sabe que ele é confiável. Consulte o tópico "Reaplicando um Modelo de Série Temporal" na página 155 para obter informações adicionais.

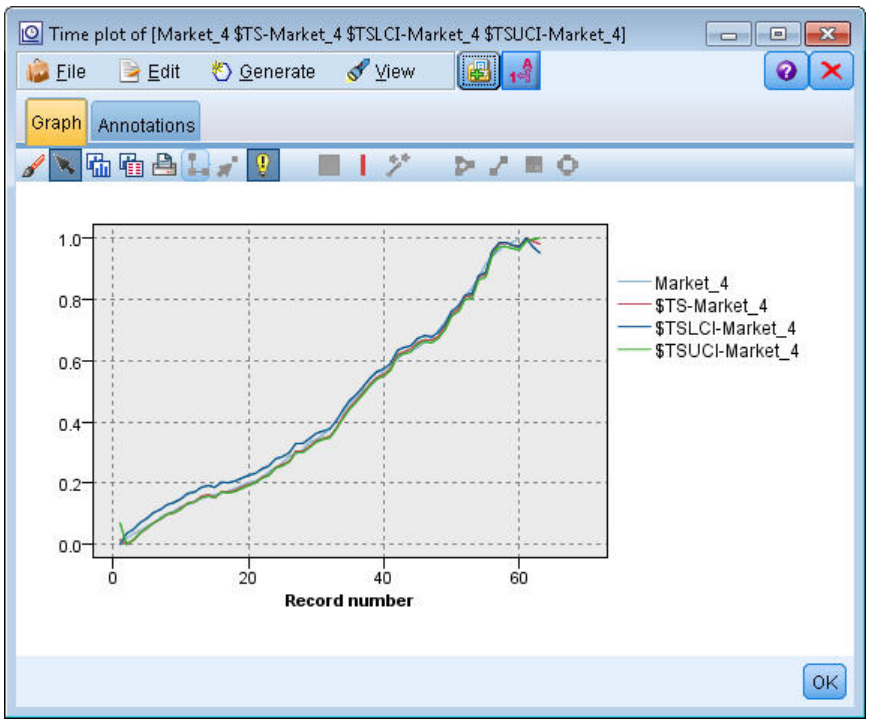


Figura 189. Gráfico de tempo com intervalo de confiança incluído

Resumo

Você aprendeu a usar o Expert Modeler para produzir previsões para várias séries temporais, e você salvou os modelos resultantes para um arquivo externo.

No próximo exemplo, você verá como transformar dados de séries temporais não padrão em um formato adequado para entrada para um nó do Time Series.

Reaplicando um Modelo de Série Temporal

Este exemplo aplica os modelos de séries temporais a partir do primeiro exemplo de série de tempo mas também pode ser usado de forma independente. Consulte o tópico "Previsão com o nó Série Temporal" na página 141 para obter informações adicionais.

Assim como no cenário original, é necessário um analista para um provedor nacional de banda larga para produzir previsões mensais de assinaturas de usuários para cada um de vários mercados locais, a fim de prever requisitos de largura de banda. Você já usou o Expert Modeler para criar modelos e para prever três meses no futuro.

O seu armazém de dados agora foi atualizado com os dados reais para o período de previsão original, portanto, você gostaria de usar esses dados para estender o horizonte de previsão por mais três meses.

Este exemplo usa o fluxo denominado *broadband_apply_models.str*, que faz referência ao arquivo de dados denominado *broadband_2.sav* Esses arquivos estão disponíveis a partir da pasta *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *broadband_apply_models.str* está na pasta *streams* .

Recuperando o Fluxo

Neste exemplo, você estará recriando um nó do Time Series a partir do modelo Time Series salvo no primeiro exemplo. Não se preocupe se você não tem um modelo salvo, nós fornecemos um na pasta *Demos* .

1. Abra o fluxo *broadband_apply_models.str* na pasta *streams* em *Demos*.

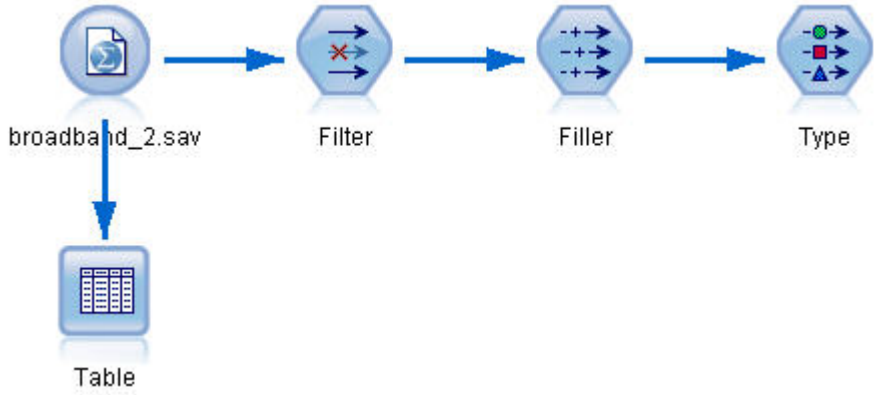
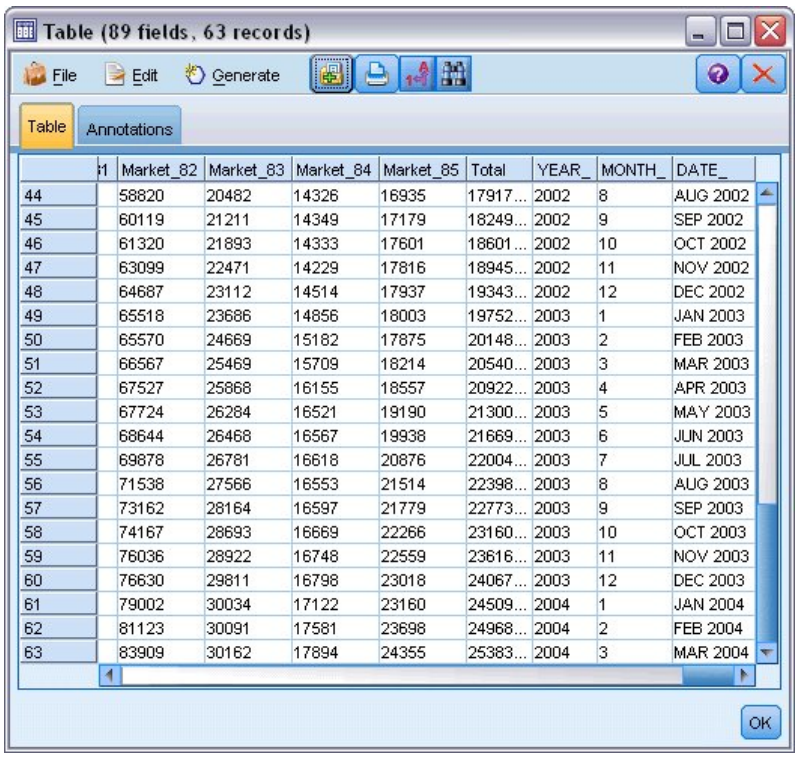


Figura 190. Abrindo o fluxo

Os dados mensais atualizados são coletados em *broadband_2.sav*.

2. Conecte um nó da Tabela ao nó de origem do Arquivo IBM SPSS Estatísticas , abra o nó da Tabela e clique em **Executar**.

   Nota: O arquivo de dados foi atualizado com os dados reais de vendas para janeiro até março de 2004, nas linhas 61 63.



	Market_82	Market_83	Market_84	Market_85	Total	YEAR_	MONTH_	DATE_
44	58820	20482	14326	16935	17917...	2002	8	AUG 2002
45	60119	21211	14349	17179	18249...	2002	9	SEP 2002
46	61320	21893	14333	17601	18601...	2002	10	OCT 2002
47	63099	22471	14229	17816	18945...	2002	11	NOV 2002
48	64687	23112	14514	17937	19343...	2002	12	DEC 2002
49	65518	23686	14856	18003	19752...	2003	1	JAN 2003
50	65570	24669	15182	17875	20148...	2003	2	FEB 2003
51	66567	25469	15709	18214	20540...	2003	3	MAR 2003
52	67527	25868	16155	18557	20922...	2003	4	APR 2003
53	67724	26284	16521	19190	21300...	2003	5	MAY 2003
54	68644	26468	16567	19938	21669...	2003	6	JUN 2003
55	69878	26781	16618	20876	22004...	2003	7	JUL 2003
56	71538	27566	16553	21514	22398...	2003	8	AUG 2003
57	73162	28164	16597	21779	22773...	2003	9	SEP 2003
58	74167	28693	16669	22266	23160...	2003	10	OCT 2003
59	76036	28922	16748	22559	23616...	2003	11	NOV 2003
60	76630	29811	16798	23018	24067...	2003	12	DEC 2003
61	79002	30034	17122	23160	24509...	2004	1	JAN 2004
62	81123	30091	17581	23698	24968...	2004	2	FEB 2004
63	83909	30162	17894	24355	25383...	2004	3	MAR 2004

Figura 191. Dados de vendas atualizados

Recuperando o modelo salvo

1. No menu IBM SPSS Modelador , escolha **Inserir** > **Nó do Arquivo** e selecione o arquivo *TSmodel.nod* na pasta *Demos* (ou use o modelo de Série Temporal salvo no primeiro exemplo de série temporal).

Este arquivo contém os modelos de séries temporais a partir do exemplo anterior. A operação de inserção coloca o nugget de modelo de Série Temporal Correspondente na tela.

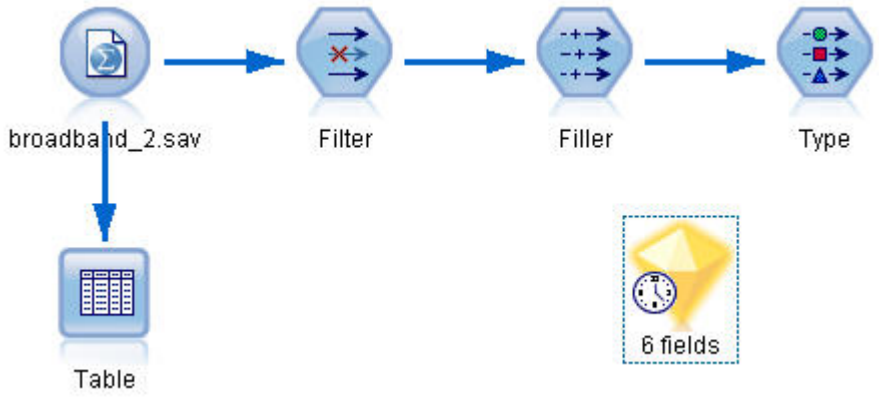


Figura 192. Adicionando o nugget modelo

Gerando um Nó de Modelagem

1. Abra o nugget do modelo Time Series e escolha **Gerar** > **Gerar Nó de Modelagem**.

Isso coloca um nó de modelagem da Série Time na tela.

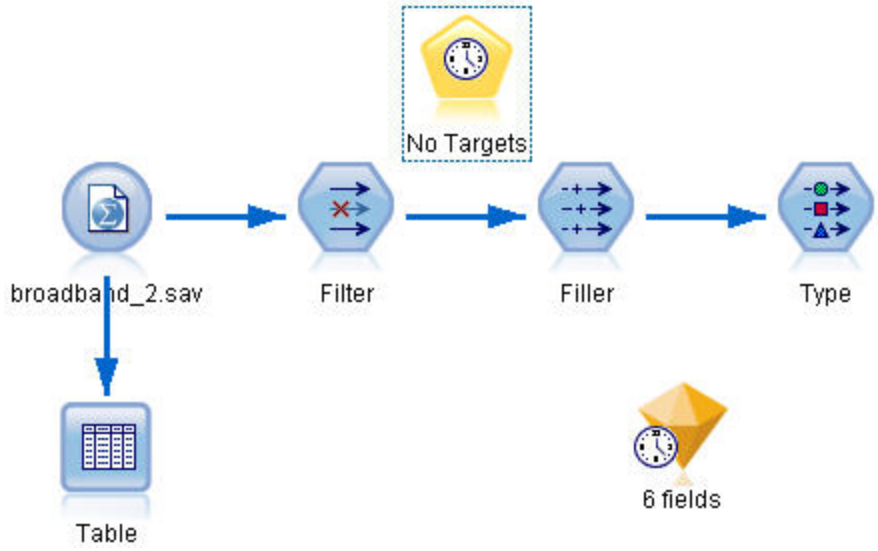


Figura 193. Gerando um nó de modelagem a partir do modelo nugget

Gerando um Novo Modelo

1. Feche o nugget do modelo Time Series e exclua-o da tela.

   O modelo antigo foi construído em 60 linhas de dados. É necessário gerar um novo modelo com base nos dados de vendas atualizados (63 linhas).
2. Conecte o nó de construção do Time Series recém-gerado para o fluxo.

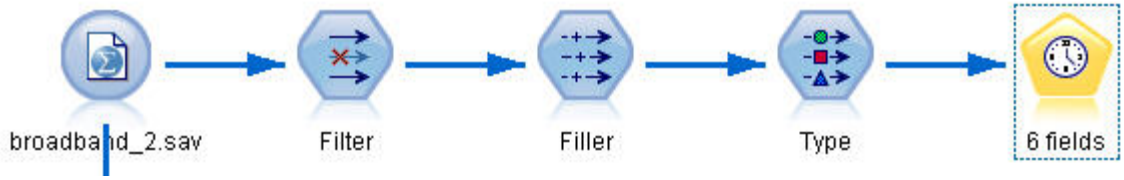


Figura 194. Conectando o nó de modelagem ao fluxo

3. Abra o nó do Time Series.

4. Na guia **Opções do Modelo** , certifica-se de que **Continuar a estimação usando modelos existentes** é verificada.

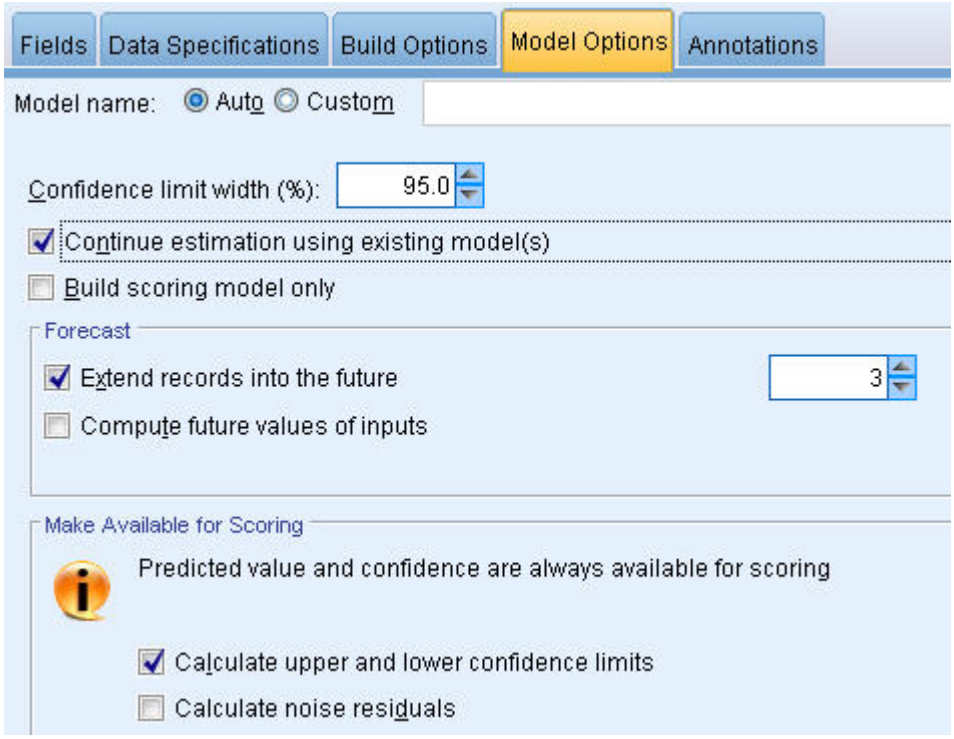


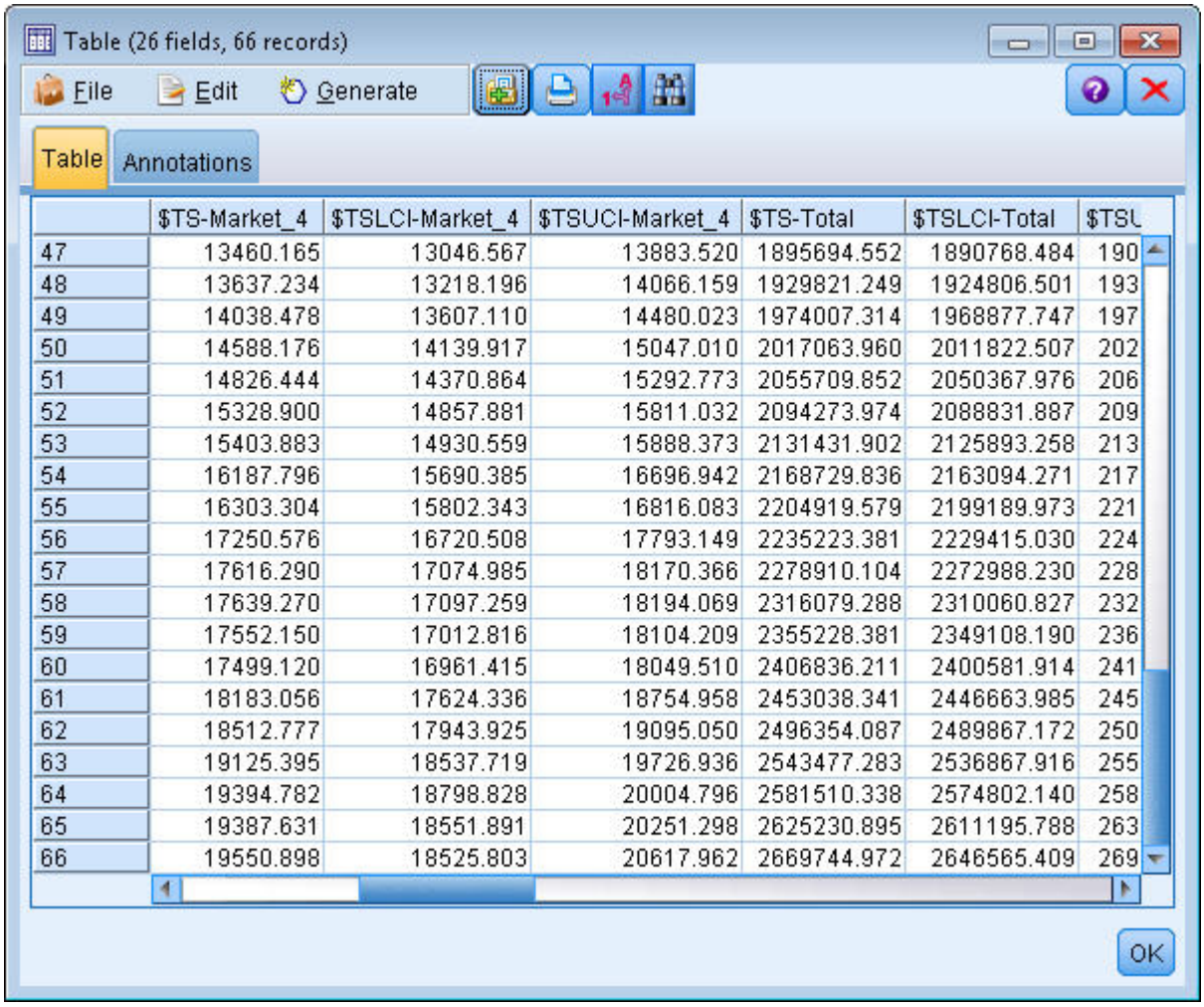
Figura 195. Como reutilizar configurações armazenadas para o modelo de série temporal

5. Certifica-se de que **Extend registros para o futuro** esteja configurado como **3**.
6. Clique em **Executar** para colocar um novo modelo nugget na tela e na paleta Models.

Examinando o Novo Modelo

1. Conecte um nó da Tabela ao novo nugget modelo da Série Time na tela.
2. Abra o nó da Tabela e clique em **Executar**.

   O novo modelo ainda prevê três meses à frente porque você está reutilizando as configurações armazenadas. No entanto, desta vez prevê abril até junho (nas linhas 64 66) porque o período de estimação agora termina em março em vez de janeiro.



	$TS-Market_4	$TSLCI-Market_4	$TSUCI-Market_4	$TS-Total	$TSLCI-Total	$TSU
47	13460.165	13046.567	13883.520	1895694.552	1890768.484	190
48	13637.234	13218.196	14066.159	1929821.249	1924806.501	193
49	14038.478	13607.110	14480.023	1974007.314	1968877.747	197
50	14588.176	14139.917	15047.010	2017063.960	2011822.507	202
51	14826.444	14370.864	15292.773	2055709.852	2050367.976	206
52	15328.900	14857.881	15811.032	2094273.974	2088831.887	209
53	15403.883	14930.559	15888.373	2131431.902	2125893.258	213
54	16187.796	15690.385	16696.942	2168729.836	2163094.271	217
55	16303.304	15802.343	16816.083	2204919.579	2199189.973	221
56	17250.576	16720.508	17793.149	2235223.381	2229415.030	224
57	17616.290	17074.985	18170.366	2278910.104	2272988.230	228
58	17639.270	17097.259	18194.069	2316079.288	2310060.827	232
59	17552.150	17012.816	18104.209	2355228.381	2349108.190	236
60	17499.120	16961.415	18049.510	2406836.211	2400581.914	241
61	18183.056	17624.336	18754.958	2453038.341	2446663.985	245
62	18512.777	17943.925	19095.050	2496354.087	2489867.172	250
63	19125.395	18537.719	19726.936	2543477.283	2536867.916	255
64	19394.782	18798.828	20004.796	2581510.338	2574802.140	258
65	19387.631	18551.891	20251.298	2625230.895	2611195.788	263
66	19550.898	18525.803	20617.962	2669744.972	2646565.409	269

Figura 196. Tabela mostrando nova previsão

3. Conecte um nó gráfico de Tempo Plot ao nugget modelo da Série Time.

   Desta vez usaremos o display de trama de tempo projetado especialmente para modelos de séries temporais.

4. Na guia Plot, configure a **etiqueta do eixo X** para **Custom**, e selecione `Date_`.
5. Para o **Plot**, escolha a opção **Selecionados modelos Time Series** .
6. A partir da lista **Série** , clique no botão seletor de campo, selecione o campo `$TS-Market_4` e clique em **OK** para adicioná-lo à lista.

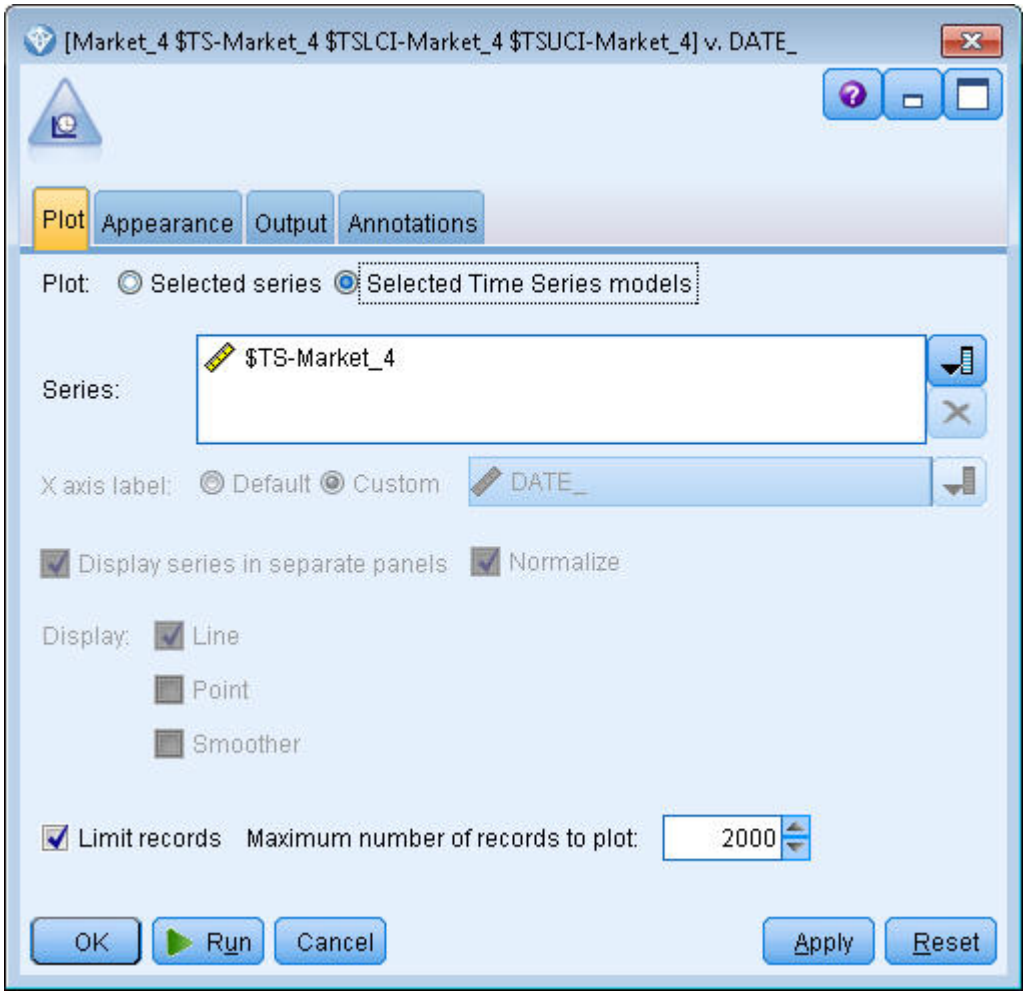


Figura 197. Especificando campos para enredo

7. Clique em **Executar** .

Agora você tem um gráfico que mostra as vendas reais para o Market_4 até março de 2004, juntamente com as vendas previstas (Previstas) e o intervalo de confiança (indicado pela área sombreada azul) até junho de 2004.

Como no primeiro exemplo, os valores previstos seguem os dados reais de perto durante todo o período de tempo, indicando, mais uma vez, que você tem um bom modelo.

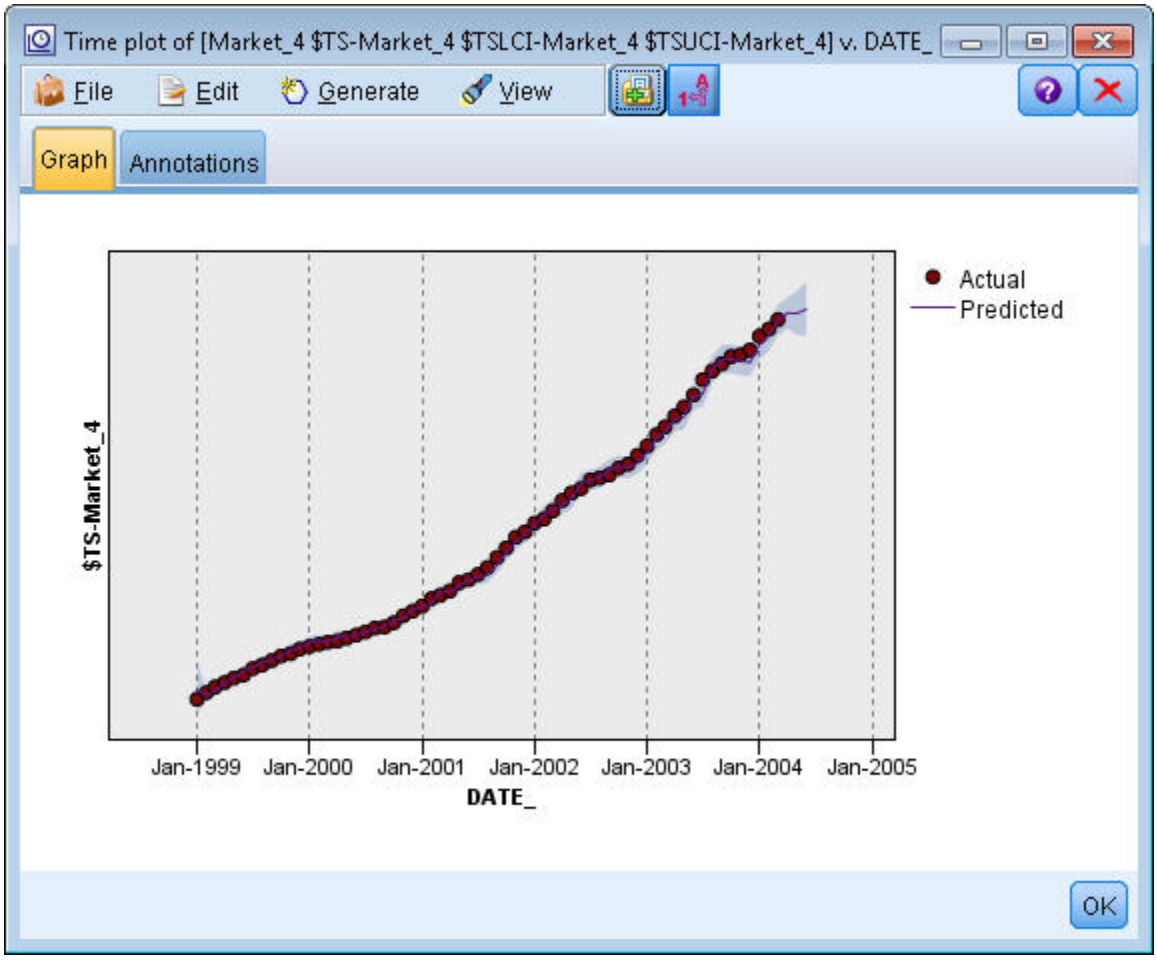


Figura 198. Previsão prorrogada para junho

Resumo

Você aprendeu como aplicar modelos salvos para estender suas previsões anteriores quando dados mais atuais se tornam disponíveis, e você fez isso sem reconstruir seus modelos. É claro que, se há razão para pensar que um modelo mudou, você deve reconstruí-o.

Capítulo 15. Previsão De Vendas Do Catálogo (Série A Tempo)

Uma empresa de catálogo está interessada em previsão de vendas mensais de sua linha de roupas masculina, com base em seus dados de vendas dos últimos 10 anos.

Este exemplo usa o fluxo denominado *catalog_forecast.str*, que faz referência ao arquivo de dados denominado *catalog_seasfac.sav* Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *catalog_forecast.str* está no diretório *streams*

Vimos em um exemplo anterior como você pode deixar o Expert Modeler decidir qual é o modelo mais apropriado para a sua série temporal. Agora é hora de ver um olhar mais atento sobre os dois métodos que estão disponíveis ao escolher um modelo você mesmo -- o smoothing exponencial e o ARIMA.

Para ajudá-lo a decidir sobre um modelo apropriado, é uma boa ideia traçar a série temporal primeiro. A inspeção visual de uma série temporal pode frequentemente ser um guia poderoso para ajudá-lo a escolher. Em particular, você precisa se perguntar:

- A série tem uma tendência geral? Em caso afirmativo, a tendência parece constante ou parece estar morrendo com o tempo?
- A série mostra sazonalidade? Em caso afirmativo, as flutuações sazonais parecem aumentar com o tempo ou parecem constantes ao longo de períodos sucessivos?

Criando o Fluxo

1. Crie um novo fluxo e inclua um nó de origem do Arquivo de Estatísticas apontando para *catalog_seasfac.sav*.

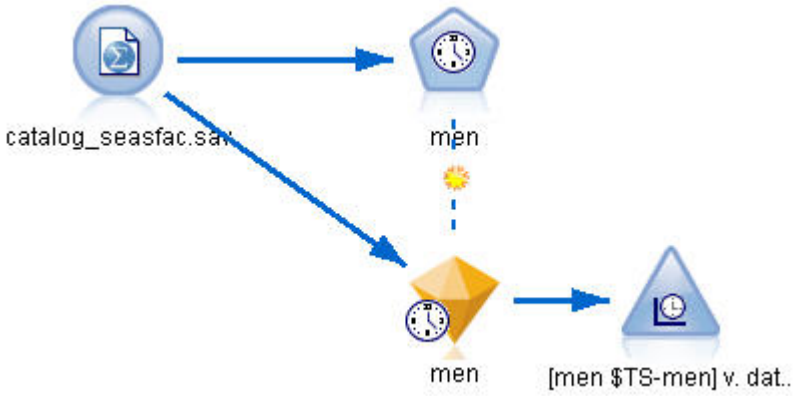


Figura 199. Previsão de vendas por catálogo

2. Abra o nó de origem do Arquivo IBM SPSS Estatísticas e selecione a guia Tipos.
3. Clique em **Valores de leitura**, em seguida, **OK**.
4. Clique na coluna **Função** para o campo men e configure a função para **Destino**.

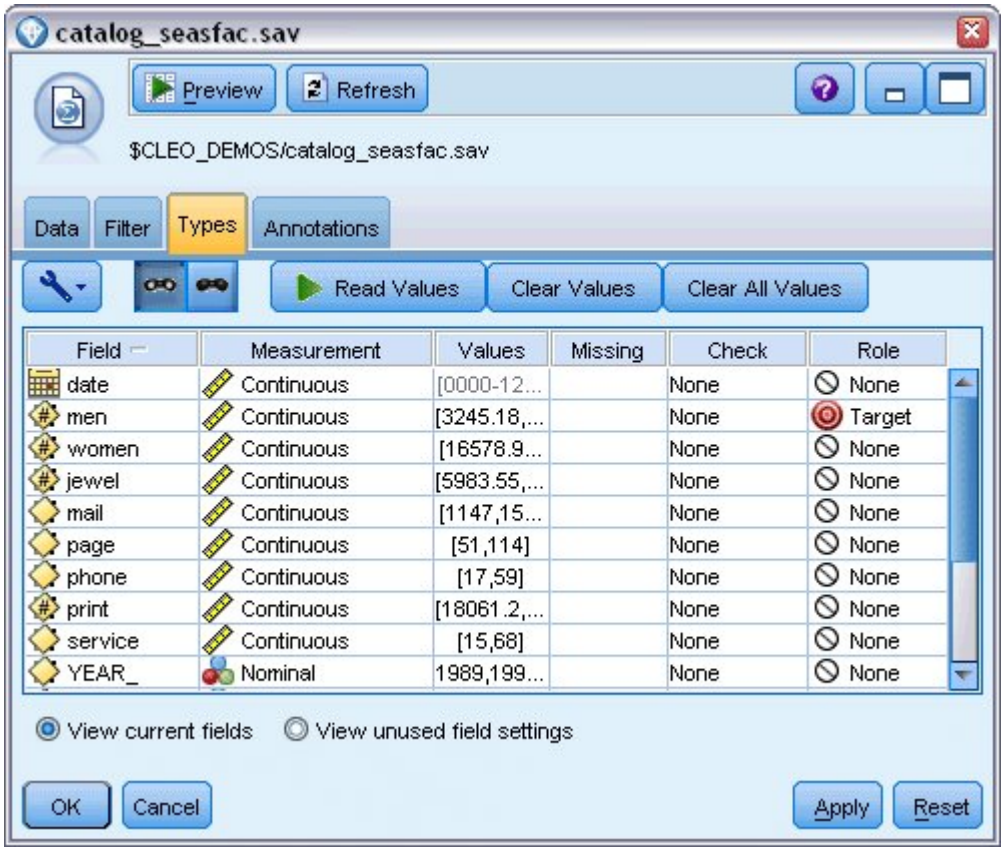


Figura 200. Especificando o campo de destino

5. Configure a função para todos os outros campos para **Nenhum**, e clique em **OK**.
6. Conecte um nó gráfico de Tempo Plot ao nó de origem do Arquivo IBM SPSS Estatísticas .
7. Abra o nó do Time Plot e, na aba Plot, adiciona `men` à lista **Series** .
8. Configure a etiqueta do eixo **X** para **Custom**, e selecione `date`.
9. Limpe a caixa de seleção **Normalizar** .

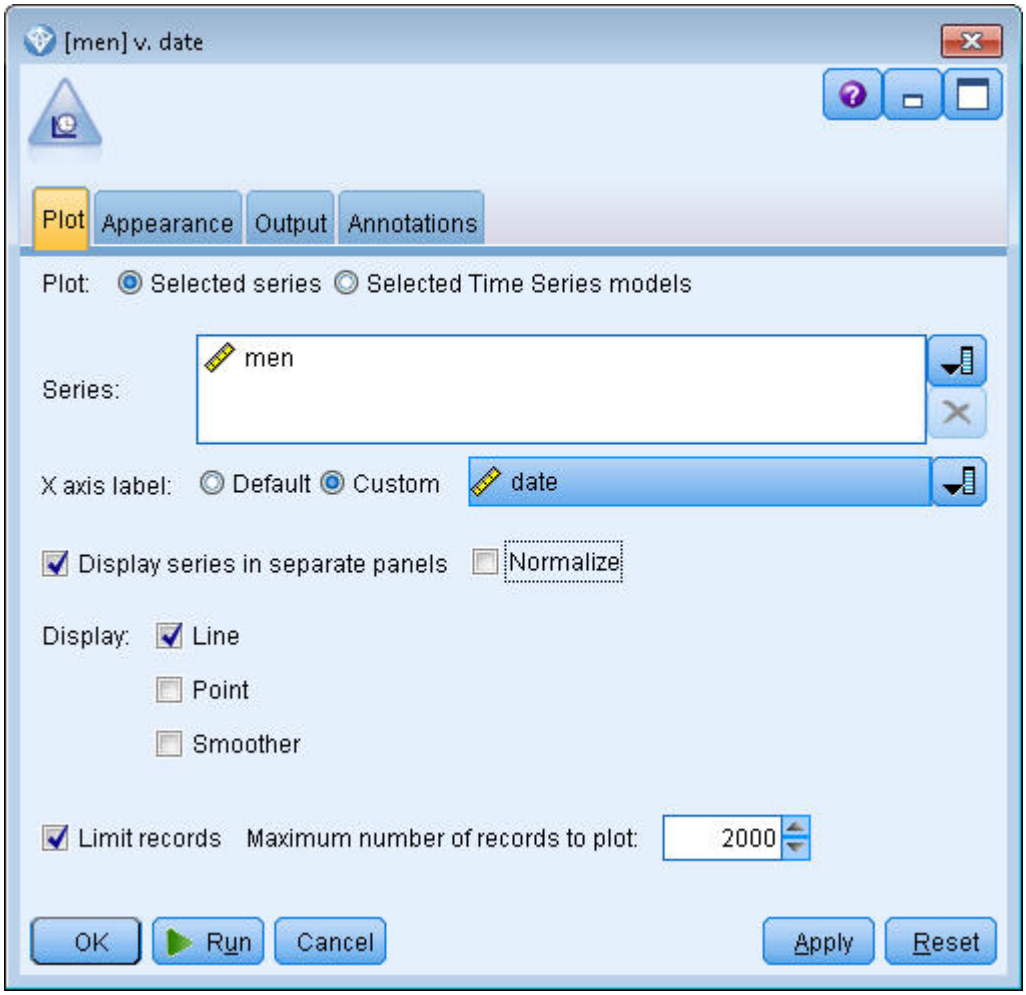


Figura 201. Plotar a série temporal

10. Clique em **Executar** .

Examinando os dados

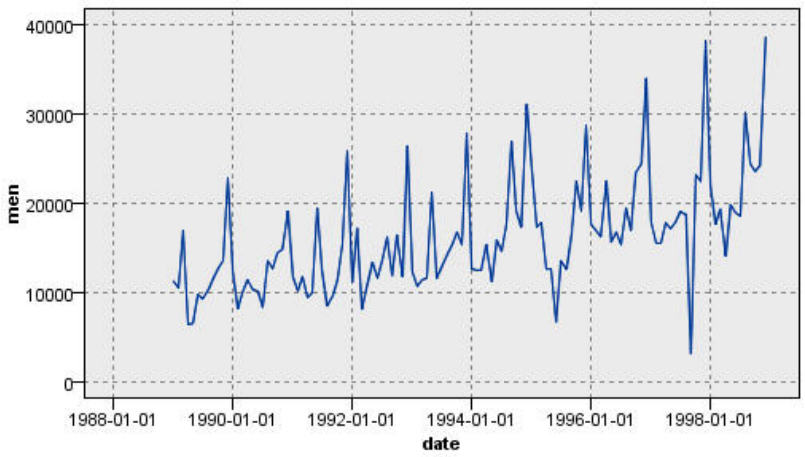


Figura 202. Vendas reais de roupas masculinas

A série mostra uma tendência crescente geral; ou seja, os valores da série tendem a aumentar com o tempo. A tendência crescente é aparentemente constante, o que indica uma tendência linear.

A série também apresenta um padrão sazonal distinto, com máximas anuais em dezembro, conforme indicado pelas retas verticais do gráfico. As variações sazonais parecem crescer com a tendência ascendente da série, o que sugere sazonalidade multiplicativa e não aditiva.

1. Clique em **OK** para fechar a trama.

Agora que identificou as características da série, você está pronto para tentar modelá-la. O método de suavização exponencial é útil para prever séries que exibem tendência, sazonalidade ou ambas. Como vimos, seus dados exibem ambas as características.

Suavização exponencial

Construir um modelo de suavização exponencial de melhor ajuste envolve determinar o tipo de modelo (se o modelo precisa incluir tendência, sazonalidade ou ambos) e, em seguida, obter os parâmetros de melhor ajuste para o modelo escolhido.

O gráfico das vendas de roupas masculinas ao longo do tempo sugeriu um modelo com um componente de tendência linear e um componente de sazonalidade multiplicativa. Isso implica um modelo de inverno. Primeiro, no entanto, exploraremos um modelo simples (sem tendência e sem sazonalidade) e, em seguida, um modelo de Holt (incorpora tendência linear, mas sem sazonalidade). Isso lhe dará prática na identificação de quando um modelo não se ajusta bem aos dados, uma habilidade essencial na construção de um modelo bem-sucedido.

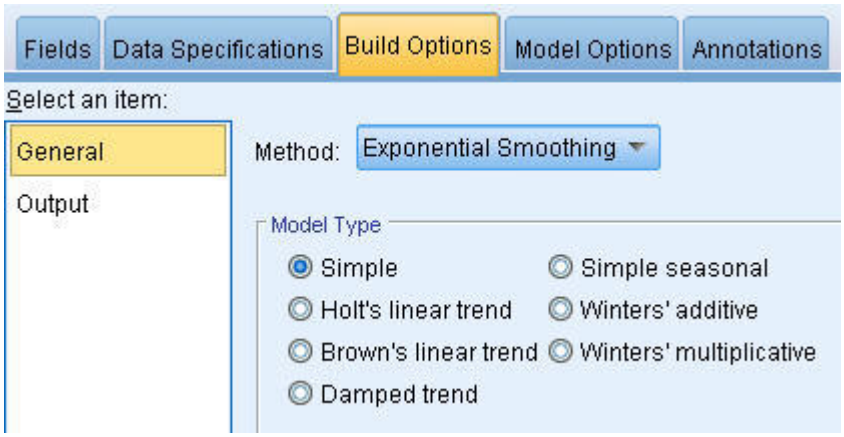


Figura 203. Especificando o Smoothing Exponencial

Começaremos com um modelo de suavização exponencial simples.

1. Inclua um nó do Time Series no fluxo e anexe-o ao nó de origem.
2. Na guia Especificações de Dados, na área de janela Observações, selecione `date` como o campo **Data / hora.**
3. Selecione `Months` como o **Intervalo de Tempo**.

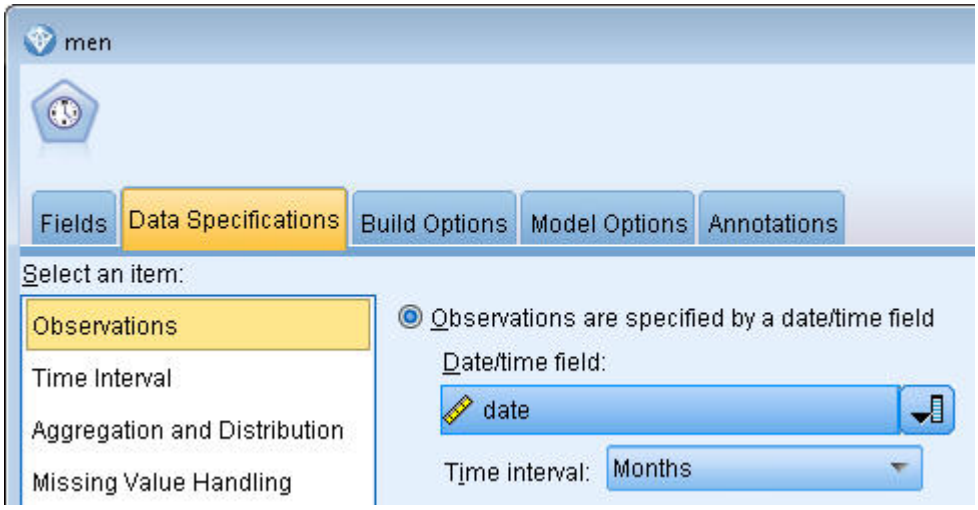


Figura 204. Configurando o intervalo de tempo

4. Na guia Opções de Construção, na área de janela Geral, configure **Método** para **Smoothing Exponential**.
5. Configure o **Tipo de Modelo** como **Simples**.

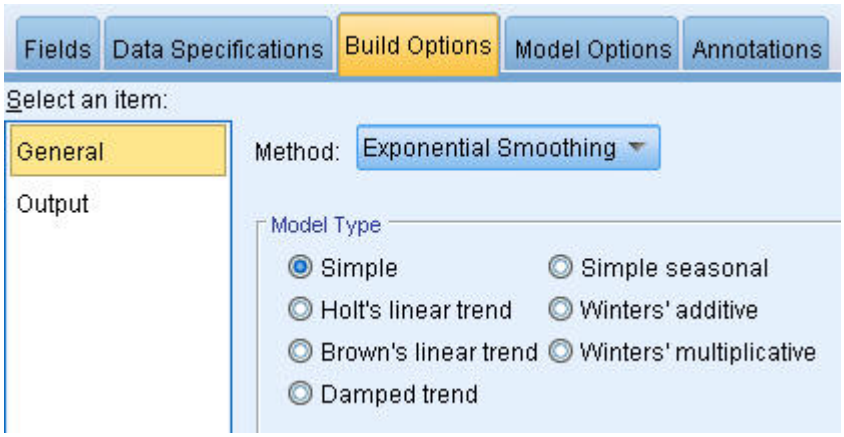


Figura 205. Configurando o método de construção do modelo

6. Clique em **Executar** para criar o nugget modelo.

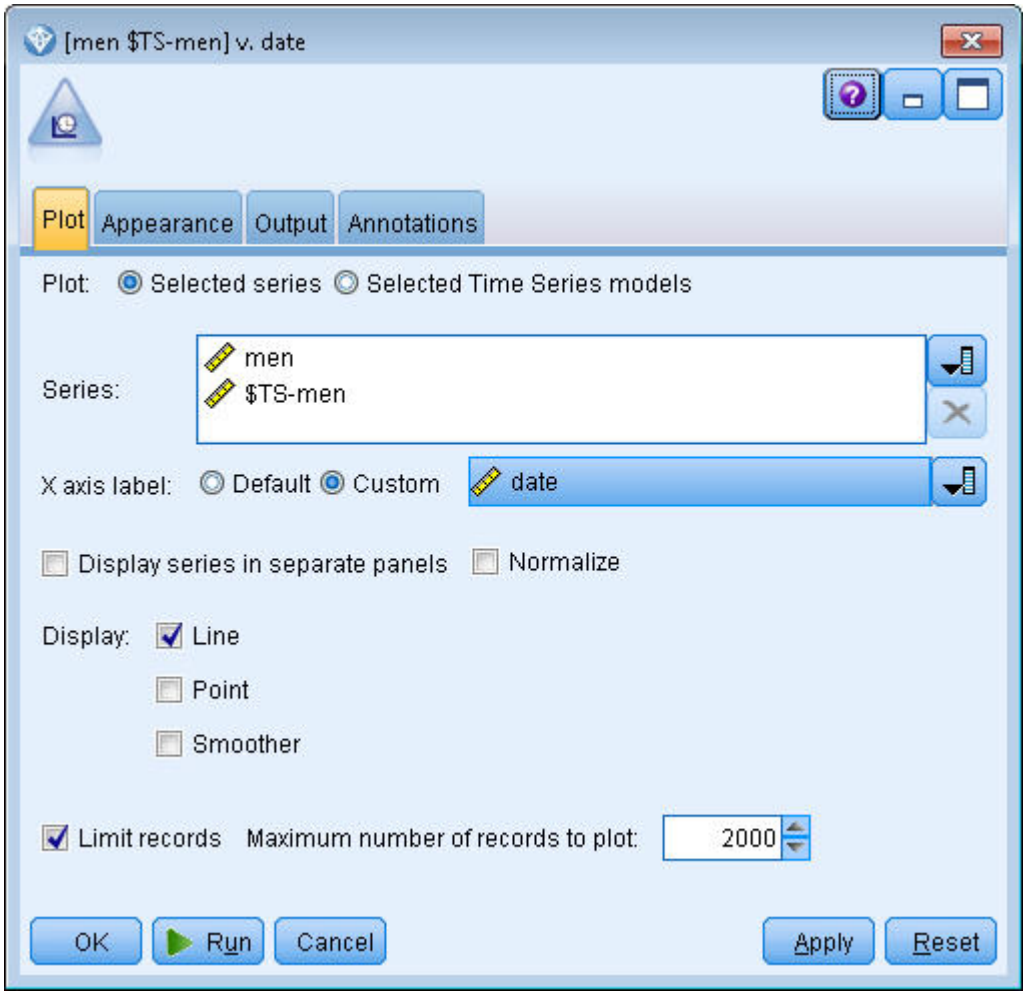


Figura 206. Plotar o modelo Time Series

7. Anexe um nó do gráfico de tempo ao nugget do modelo.
8. Na guia **Plot** , inclua `men` e `$TS-men` na lista **Série** .
9. Configure o **Rótulo do eixo X** como **Customizado**e selecione `date`.
10. Limpe a **Série Exibir em painéis separados** e **Normalize** caixas de seleção.
11. Clique em **Executar** .

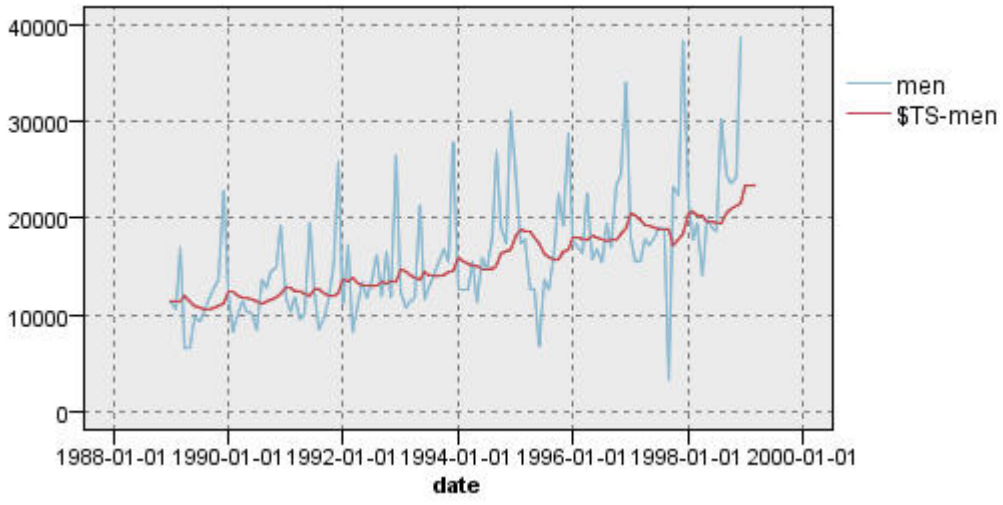


Figura 207. Modelo de suavização exponencial simples

A parcela **men** representa os dados reais, enquanto **$TS-men** denota o modelo de série temporal.

Embora o modelo simples faça, de fato, uma tendência de alta gradual (e bastante ponderosa), não leva em consideração a sazonalidade. É possível rejeitar este modelo com segurança.

12. Clique em **OK** para fechar a janela de trama de tempo.

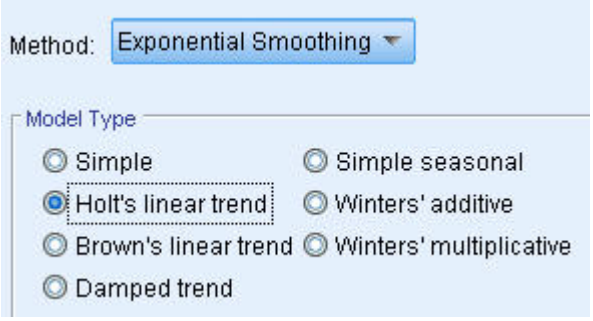


Figura 208. Seleção do modelo de Holt

Vamos tentar o modelo linear de Holt. Isso deve pelo menos modelar a tendência melhor do que o modelo simples, embora também seja improvável que capture a sazonalidade.

13. Reabra o nó do Time Series.
14. Na guia Opções de Construção, na pane Geral, com **Smoothing Exponential** ainda selecionada como o **Método**, selecione **Holts tendência linear** como o **Tipo de Modelo**.
15. Clique em **Executar** para recriar o nugget modelo.
16. Re-abra o nó do Time Plot e clique em **Executar**.

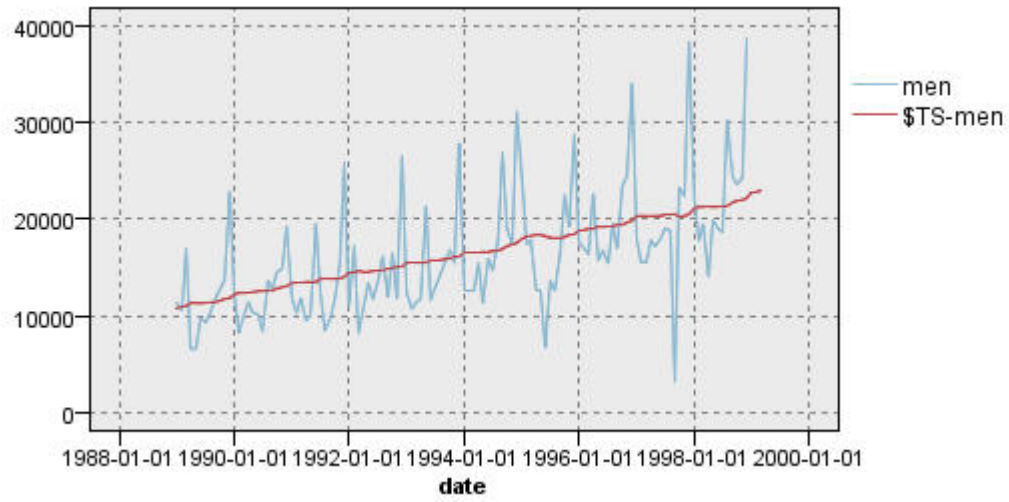


Figura 209. Modelo de tendência linear de Holt

O modelo de Holt exibe uma tendência de alta defumada do que o modelo simples mas ainda não leva em conta a sazonalidade, então você pode descartar este também.

17. Fechar a janela de trama de tempo.

Você deve se lembrar que o gráfico inicial das vendas de roupas masculinas ao longo do tempo sugeria um modelo que incorporava uma tendência linear e sazonalidade multiplicativa. Um candidato mais adequado, portanto, pode ser o modelo de inverno.

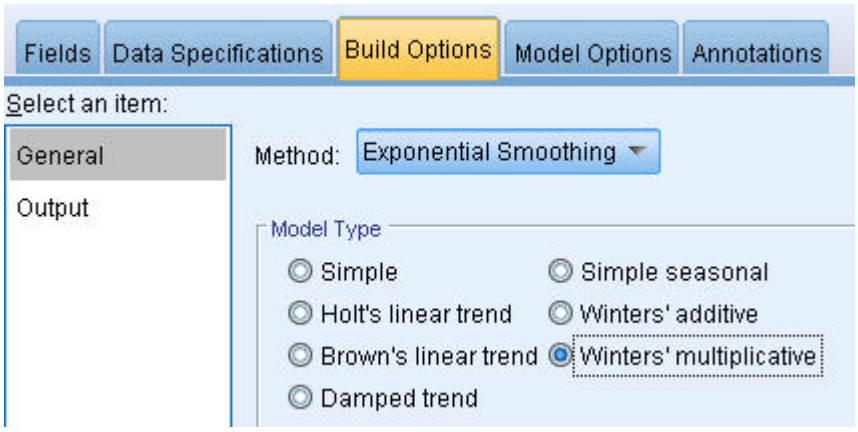


Figura 210. Seleção do modelo de Winters

18. Reabra o nó do Time Series.
19. Na guia Opções de Construção, na pane Geral, com **Smoothing Exponential** ainda selecionada como o **Método**, selecione **Winters ' multiplicative** como o **Tipo de Modelo**.
20. Clique em **Executar** para recriar o nugget modelo.
21. Abra o nó do Time Plot e clique em **Executar**.

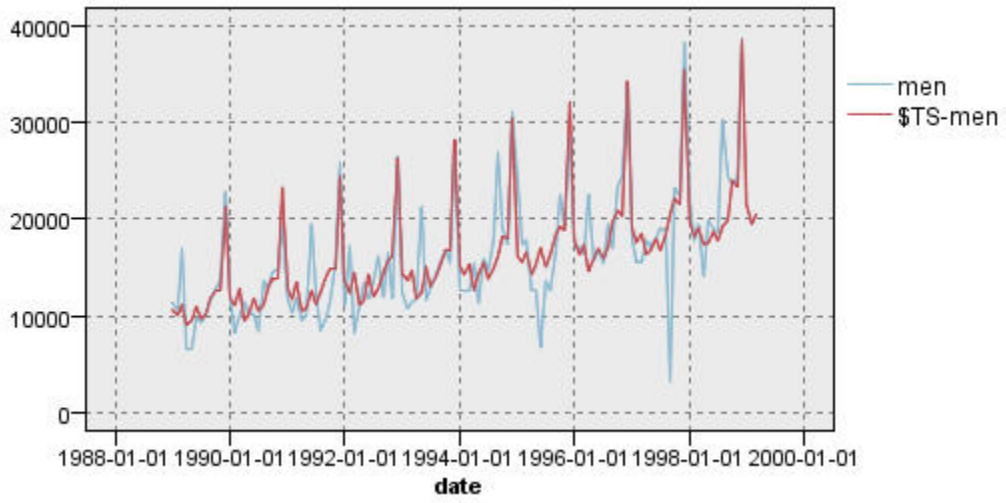


Figura 211. Modelo multiplicativo de inverno

Isso parece melhor; o modelo reflete tanto a tendência quanto a sazonalidade dos dados.

O conjunto de dados cobre um período de 10 anos e inclui 10 picos sazonais que ocorrem em dezembro de cada ano. Os 10 picos presentes nos resultados previstos correspondem bem aos 10 picos anuais nos dados reais.

No entanto, os resultados também enfatizam as limitações do procedimento de suavização exponencial. Olhando tanto para os picos para cima quanto para baixo, há uma estrutura significativa que não é contabilizados.

Se você está primariamente interessado em modelar uma tendência de longo prazo com variação sazonal, então o smoothing exponencial pode ser uma boa escolha. Para modelar uma estrutura mais complexa como esta, precisamos considerar o uso do procedimento ARIMA.

ARIMA

Com o procedimento ARIMA você pode criar um modelo de movimento-média móvel (ARIMA) autoregressivo que é adequado para modelagem finamente ajustada de séries temporais. Os modelos

ARIMA fornecem métodos mais sofisticados para modelar tendências e componentes sazonais do que os modelos de suavização exponencial e têm o benefício adicional de poder incluir variáveis preditoras no modelo.

Continuando o exemplo da empresa de catálogo que deseja desenvolver um modelo de previsão, vimos como a empresa coletou dados sobre as vendas mensais de roupas masculinas junto com várias séries que podem ser usadas para explicar algumas das variações nas vendas. Possíveis preditores incluem o número de catálogos enviados e o número de páginas no catálogo, o número de linhas telefônicas abertas para pedidos, a quantia gasta em publicidade impressa e o número de representantes de atendimento ao cliente.

Algum desses preditores é útil para previsões? Um modelo com preditores é realmente melhor do que outro sem? Usando o procedimento ARIMA, podemos criar um modelo de previsão com preditores, e ver se há uma diferença significativa na capacidade preditiva sobre o modelo de defumação exponencial sem nenhum preditor.

Com o método ARIMA, é possível ajustar o modelo especificando ordens de autorregressão, diferenciação e média móvel, bem como contrapartidas sazonais para esses componentes. Determinar os melhores valores para esses componentes manualmente pode ser um processo demorado que envolve muitas tentativas e erros, portanto, para este exemplo, vamos deixar o Modelador Especialista escolher um modelo ARIMA para nós.

Tentaremos construir um modelo melhor tratando algumas das outras variáveis no conjunto de dados como variáveis preditoras. Os que parecem mais úteis para incluir como preditores são o número de catálogos enviados (`mail`), o número de páginas no catálogo (`page`), o número de linhas telefônicas abertas para pedidos (`phone`), a quantia gasta em publicidade impressa (`print`) e o número de representantes de atendimento ao cliente (`service`).

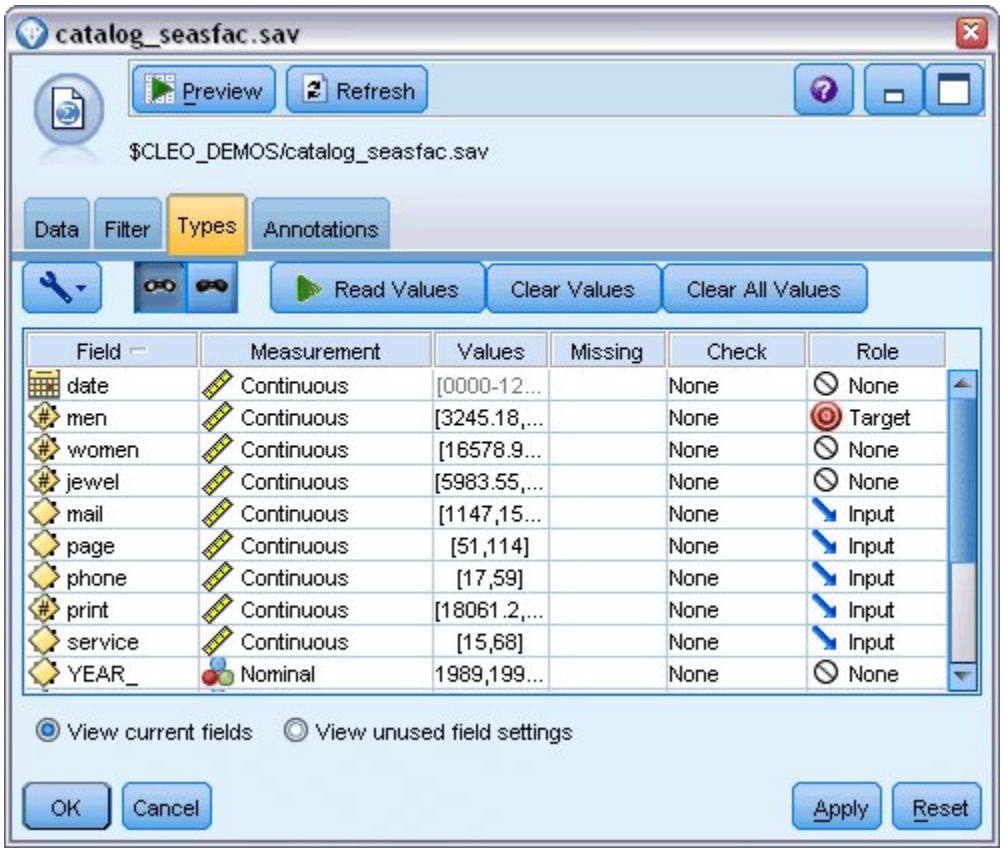


Figura 212. Configurando os campos do preditor

1. Abra o nó de origem do Arquivo IBM SPSS Estatísticas .
2. Na guia Tipos, configure **Função** para `mail`, `page`, `phone`, `printe` `service` para **Entrada**.

3. Assegure-se de que a função para men esteja configurada como **Destino** e que todos os campos restantes estejam configurados para **Nenhum**
4. Clique em **OK**.
5. Abra o nó do Time Series.
6. Na guia Opções de Construção, na área de janela Geral, configure **Método** para **Expert Modeler**.
7. Selecione a opção **Somente modelos ARIMA** e certise-se de que **Expert Modeler considera modelos sazonais** é verificado.

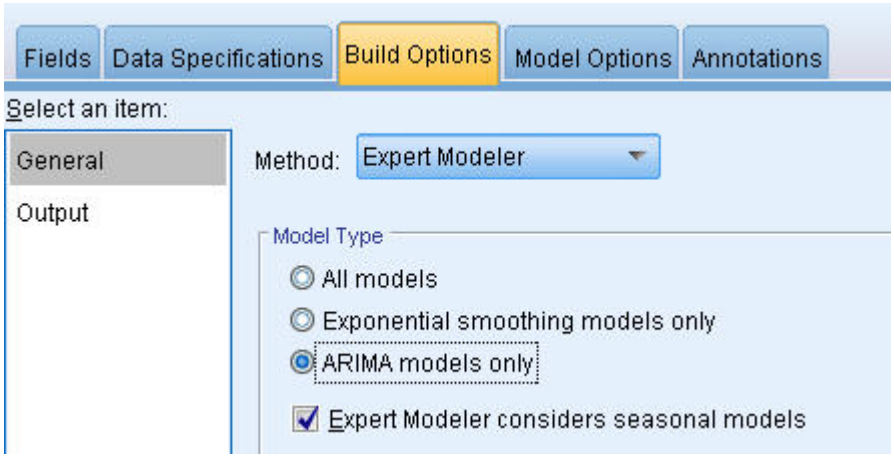


Figura 213. Escolhendo apenas modelos ARIMA

8. Clique em **Executar** para recriar o nugget modelo.
9. Abra o nugget modelo.

   Na guia Output, na coluna esquerda, selecione as **Informações do Modelo**. Observe como o Modelador Especialista escolheu apenas dois dos cinco preditores especificados como sendo significativos para o modelo.

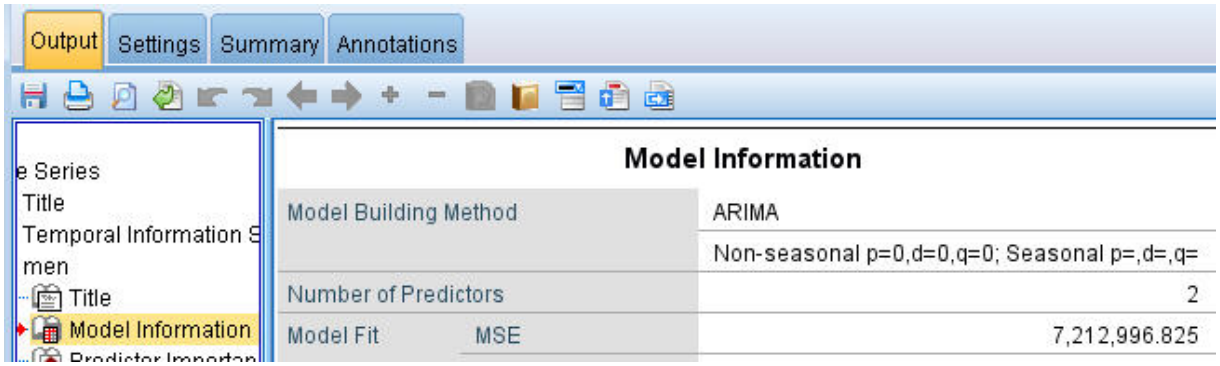


Figura 214. O Modelador Especialista escolhe dois preditores

10. Clique em **OK** para fechar o nugget modelo.
11. Abra o nó do Time Plot e clique em **Executar**.

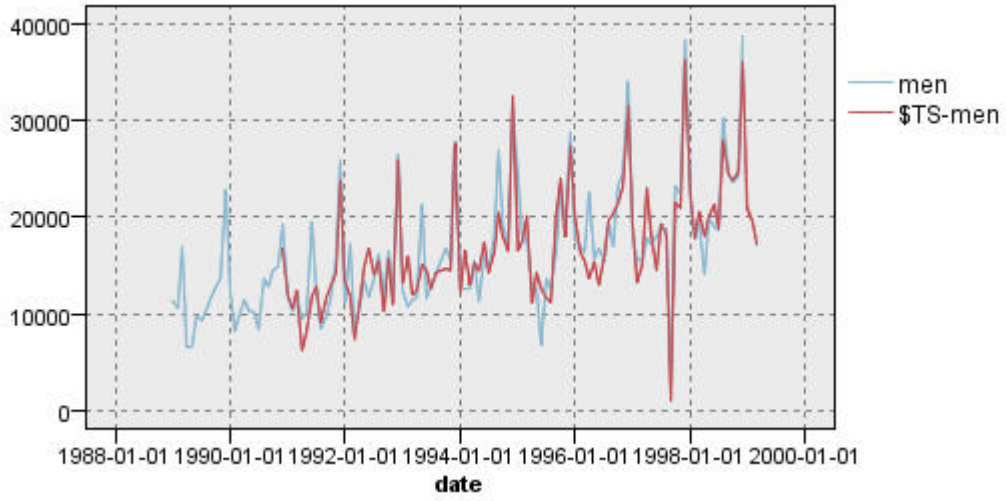


Figura 215. Modelo ARIMA com preditores especificados

Este modelo melhora o anterior, capturando também o grande pico para baixo, tornando-o a melhor até agora.

Poderíamos tentar refinar o modelo ainda mais, mas quaisquer melhorias a partir deste ponto provavelmente serão mínimas. Estabelecemos que o modelo ARIMA com preditores é preferível, então vamos usar o modelo que acabamos de construir. Para os fins deste exemplo, projetaremos as vendas para o próximo ano.

12. Clique em **OK** para fechar a janela de trama de tempo.
13. Abra o nó da Série Time e selecione a guia Opções do Modelo.
14. Selecione a caixa de seleção **Extend records no futuro** e configure seu valor para 12.
15. Selecione a caixa de opção **Compute futuro valores de entradas** .
16. Clique em **Executar** para recriar o nugget modelo.
17. Abra o nó do Time Plot e clique em **Executar**.

A previsão para 1999 parece boa; como esperado, há um retorno aos níveis normais de vendas seguindo o pico de dezembro, e uma tendência de alta constante na segunda metade do ano, com vendas em geral acima das referentes ao ano anterior.

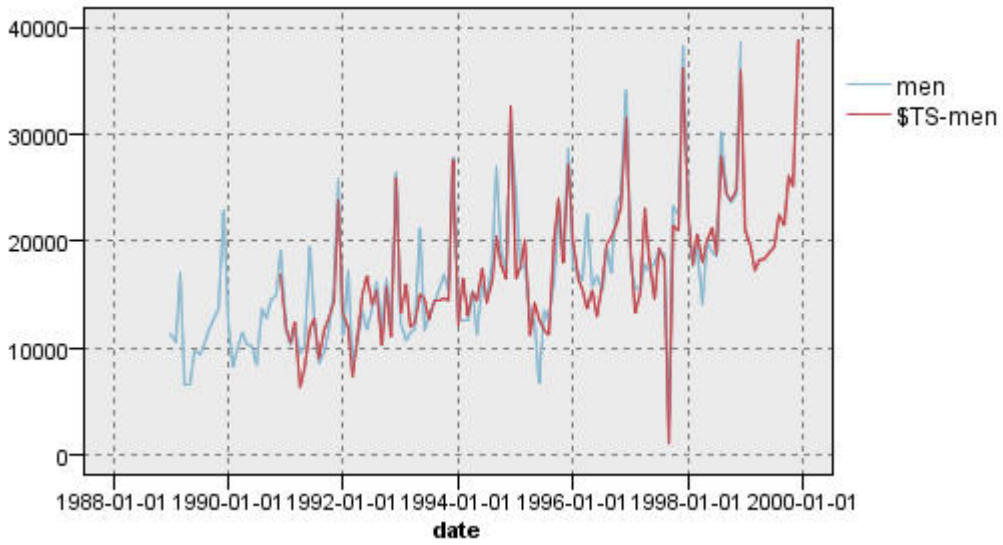


Figura 216. Previsão de vendas estendida por 12 meses

Resumo

Você modelou com sucesso uma série temporal complexa, incorporando não apenas uma tendência ascendente mas também sazonal e outras variações. Você também viu como, através de tentativa e erro, você pode ficar mais perto e mais perto de um modelo preciso, que você então usou para prever vendas futuras.

Na prática, você precisaria reaplicar o modelo à medida que seus dados reais de vendas são atualizados -- por exemplo, a cada mês ou a cada trimestre -- e produzir previsões atualizadas. Consulte o tópico “Reaplicando um Modelo de Série Temporal” na página 155 para obter informações adicionais.

Capítulo 16. Fazendo ofertas aos clientes (autoaprendizado)

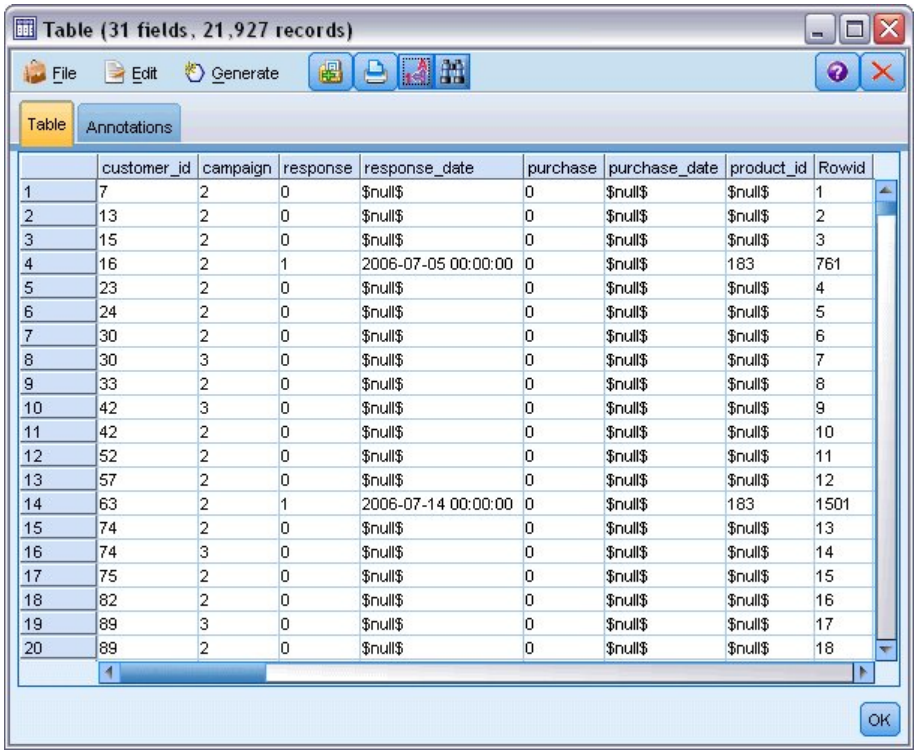
O nó Modelo de Resposta de autoaprendizado (SLRM) gera e permite a atualização de um modelo que permite prever quais ofertas são mais adequadas para os clientes e a probabilidade de as ofertas serem aceitas. Esses tipos de modelos são mais benéficos no gerenciamento de relacionamento com o cliente, como aplicativos de marketing ou call centers.

Este exemplo é baseado em uma empresa bancária fictícia. O departamento de marketing deseja obter resultados mais rentáveis em campanhas futuras, combinando a oferta certa de serviços financeiros para cada cliente. Especificamente, o exemplo usa um modelo de Resposta de Autoaprendizado para identificar as características dos clientes que têm maior probabilidade de responder favoravelmente com base em ofertas e respostas anteriores e para promover a melhor oferta atual com base nos resultados.

Este exemplo usa o fluxo *pm_selflearn.str*, que faz referência aos arquivos de dados *pm_customer_train1.sav*, *pm_customer_train2.save* *pm_customer_train3.sav* Esses arquivos estão disponíveis a partir da pasta *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *pm_selflearn.str* está na pasta *streams*

Dados Existentes

A empresa tem dados históricos rastreando as ofertas feitas aos clientes em campanhas passadas, juntamente com as respostas a essas ofertas. Esses dados também incluem informações demográficas e financeiras que podem ser usadas para prever taxas de resposta para diferentes clientes.



	customer_id	campaign	response	response_date	purchase	purchase_date	product_id	Rowid
1	7	2	0	$null$	0	$null$	$null$	1
2	13	2	0	$null$	0	$null$	$null$	2
3	15	2	0	$null$	0	$null$	$null$	3
4	16	2	1	2006-07-05 00:00:00	0	$null$	183	761
5	23	2	0	$null$	0	$null$	$null$	4
6	24	2	0	$null$	0	$null$	$null$	5
7	30	2	0	$null$	0	$null$	$null$	6
8	30	3	0	$null$	0	$null$	$null$	7
9	33	2	0	$null$	0	$null$	$null$	8
10	42	3	0	$null$	0	$null$	$null$	9
11	42	2	0	$null$	0	$null$	$null$	10
12	52	2	0	$null$	0	$null$	$null$	11
13	57	2	0	$null$	0	$null$	$null$	12
14	63	2	1	2006-07-14 00:00:00	0	$null$	183	1501
15	74	2	0	$null$	0	$null$	$null$	13
16	74	3	0	$null$	0	$null$	$null$	14
17	75	2	0	$null$	0	$null$	$null$	15
18	82	2	0	$null$	0	$null$	$null$	16
19	89	3	0	$null$	0	$null$	$null$	17
20	89	2	0	$null$	0	$null$	$null$	18

Figura 217. Respostas a ofertas anteriores

Construindo o Fluxo

1. Adicione um nó de origem do Arquivo de Estatísticas apontando para *pm_customer_train1.sav*, localizado na pasta *Demos* de sua instalação IBM SPSS Modelador .

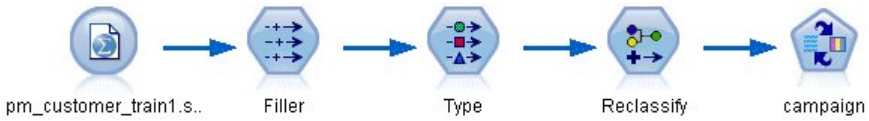


Figura 218. Fluxo de amostra SLRM

2. Inclua um nó Preenchimento e selecione `campaign` como o campo Preenchimento.
3. Selecione um tipo Replace de **Sempre**.
4. Na Substituição por caixa de texto, insira `to_string(campaign)` e clique em **OK**.

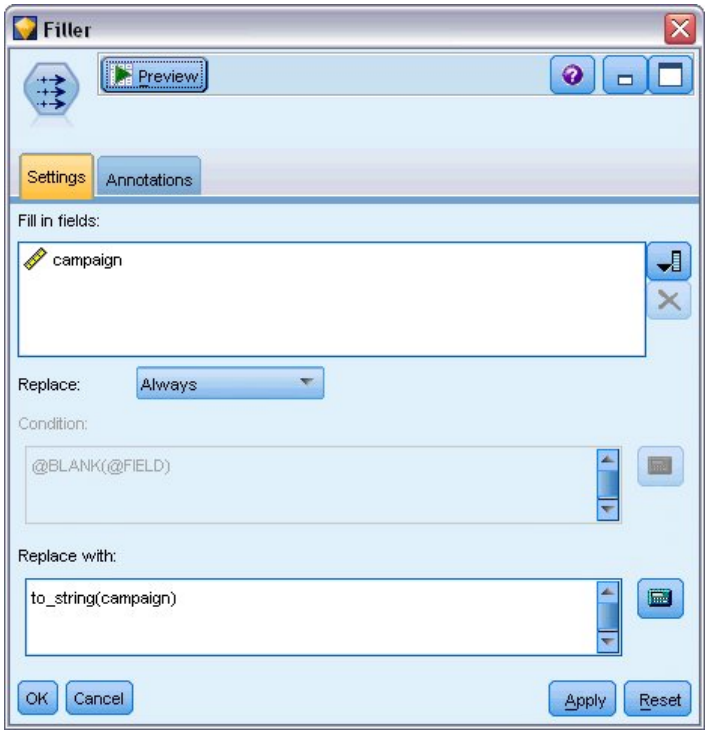


Figura 219. Derivar um campo da campanha

5. Inclua um nó Type, e configure a *Função* para **Nenhum** para o *customer_id*, *response_date*, *comprador*, Campos de *product_id*, *Rowide X_random* .

Figura 220. Alterando as configurações do nó do Tipo

6. Configure a *Função* para **Destino** para os campos *campanha* e *resposta*. Esses são os campos nos quais você deseja basear suas previsões.

   Configureao **Medição** para **Sinalizador** para o campo de *resposta*.

7. Clique em **Valores de leitura**, em seguida, **OK**.

   Como os dados do campo da campanha mostram como uma lista de números (1, 2, 3 e 4), é possível reclassificar os campos para ter títulos mais significativos.

8. Adicio um nó Reclassify ao nó Type.
9. No campo **Reclassify into** , selecione **Campo Existente**.
10. Na lista **campo Reclassify** , selecione **campanha**.
11. Clique no botão **Obter** ; os valores da campanha são adicionados à coluna *Valor Original* .
12. Na coluna *NOVO VALOR*, insira os nomes de campanha a seguir nas primeiras quatro linhas:

    - **Hipoteca**
    - **Empréstimo de carro**
    - **Poupança**
    - **Pensão**

13. Clique em **OK**.

Figura 221. Reclassifique os nomes da campanha

14. Conecte um nó de modelagem SLRM ao nó reclassificar. Na guia Campos, selecione **campanha** para o campo Destino, e **resposta** para o campo de resposta Destino.

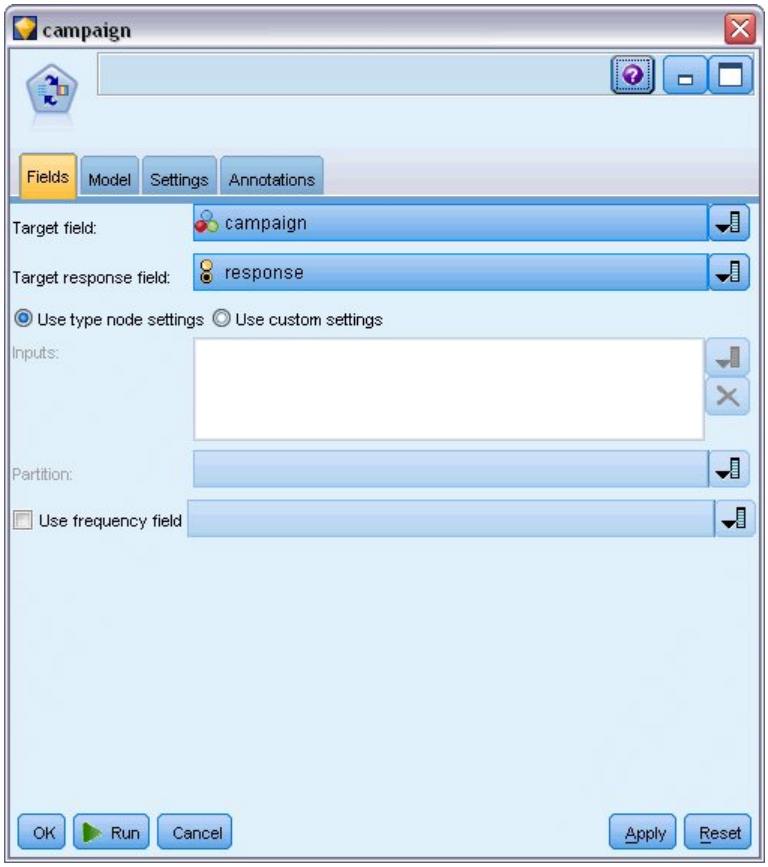


Figura 222. Selecione a resposta de destino e destino

15. Na guia Configurações, no número máximo de previsões por registro de campo, reduza o número para 2.

    Isso significa que, para cada cliente, haverá duas ofertas identificadas que têm a maior probabilidade de serem aceitas.

16. Certifica-se de que **Tenha em conta a confiabilidade do modelo** é selecionado e clique em **Executar**.

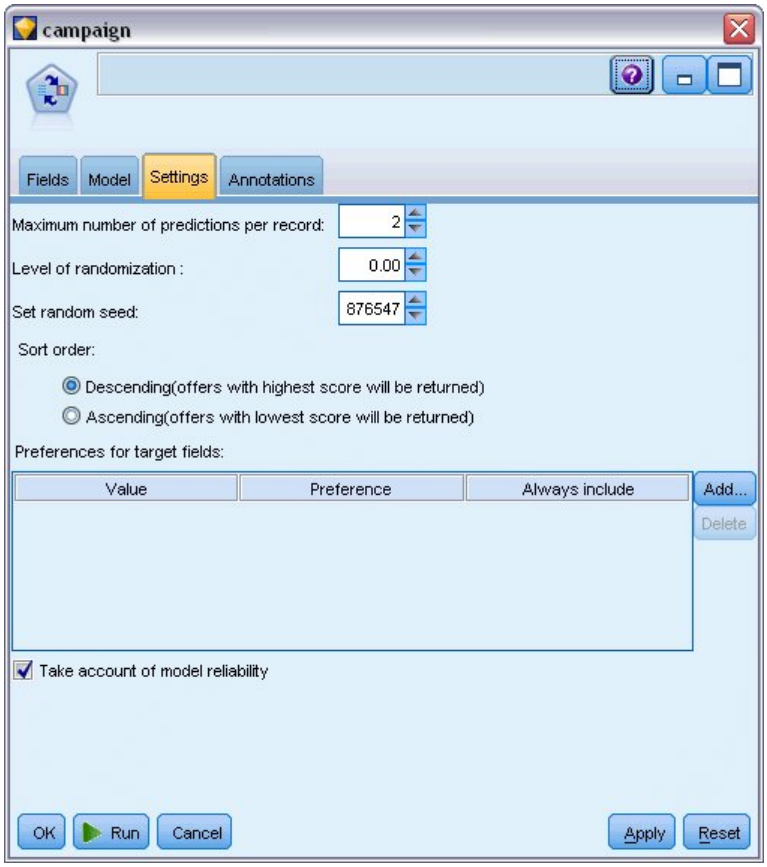


Figura 223. Configurações do nó SLRM

Navegando no modelo

1. Abra o nugget modelo. A guia Modelo mostra inicialmente a precisão estimada das previsões para cada oferta e a importância relativa de cada preditor na estimativa do modelo.

   Para exibir a correlação de cada preditor com a variável de destino, escolha **Associação com Resposta** a partir da lista **Visualizar** no painel da direita.
2. Para alternar entre cada uma das quatro ofertas para as quais há previsões, selecione a oferta necessária a partir da lista **Visualizar** no painel esquerdo.

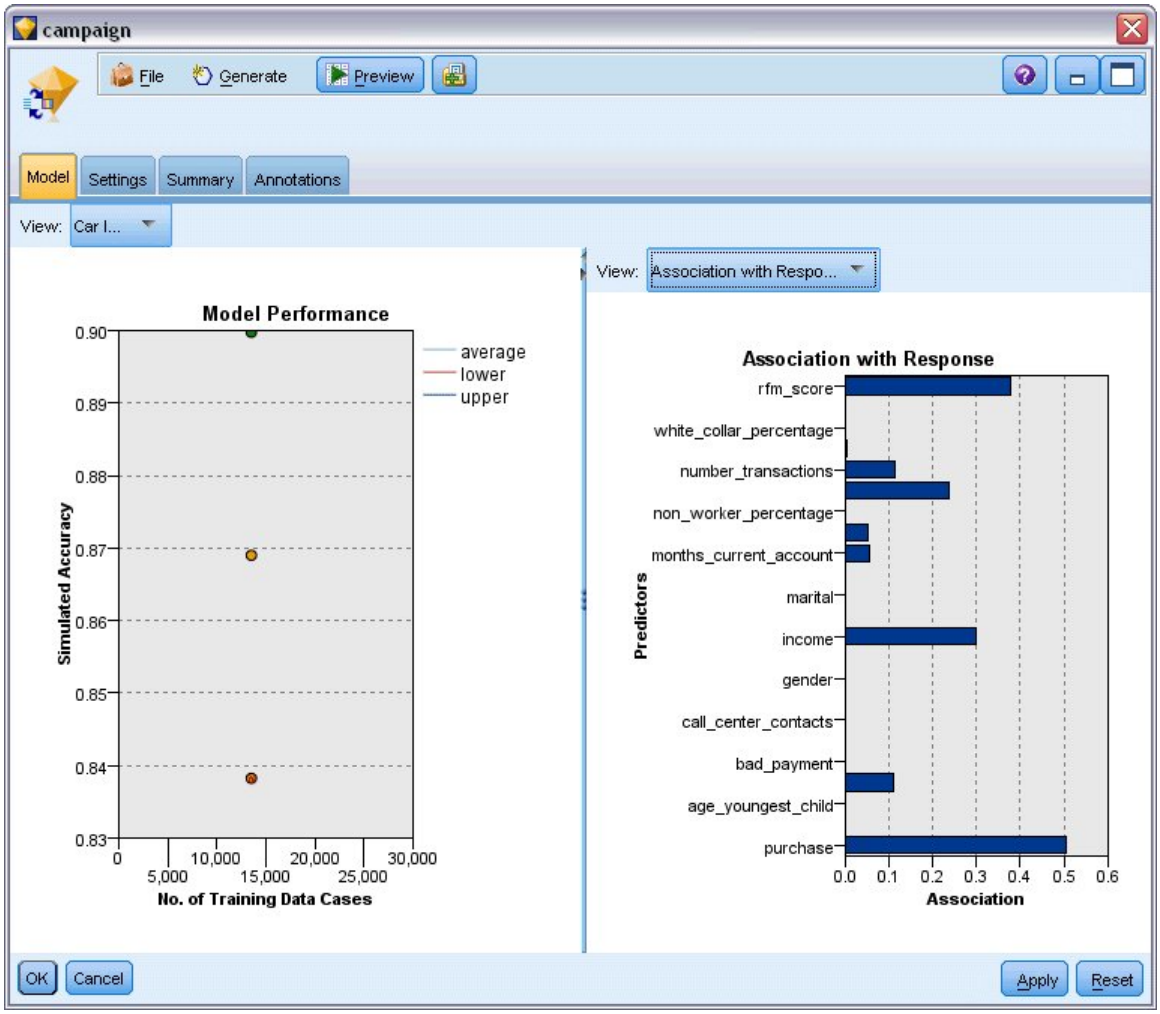


Figura 224. Nugget de modelo SLRM

3. Feche a janela de nugget modelo.
4. Na tela do fluxo, desconecte o nó de origem do Arquivo IBM SPSS Estatísticas apontando para *pm_customer_train1.sav*.
5. Adicione um nó de origem do Arquivo de Estatísticas apontando para *pm_customer_train2.sav*, localizado na pasta *Demos* de sua instalação IBM SPSS Modelador , e conecte-o ao nó Filler.

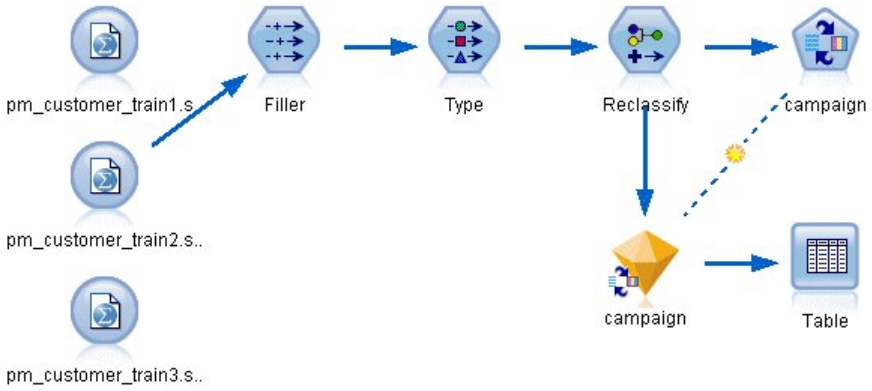


Figura 225. Conectando a segunda origem de dados ao fluxo SLRM

6. Na guia Modelo do nó SLRM, selecione **Continuar treinando modelo existente**.

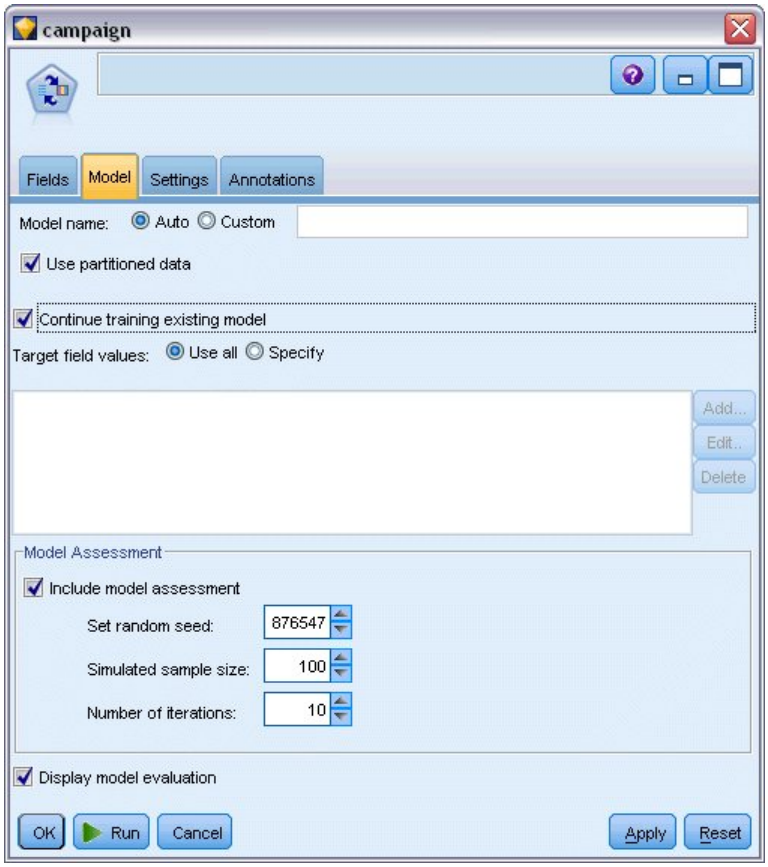


Figura 226. Continuar modelo de treinamento

7. Clique em **Executar** para recriar o nugget modelo. Para visualizar seus detalhes, dê um duplo clique no nugget sobre a tela.

   A guia Modelo agora mostra as estimativas revisadas sobre a precisão das previsões para cada oferta.

8. Adicione um nó de origem do Arquivo de Estatísticas apontando para *pm_customer_train3.sav*, localizado na pasta *Demos* de sua instalação IBM SPSS Modelador , e conecte-o ao nó Filler.

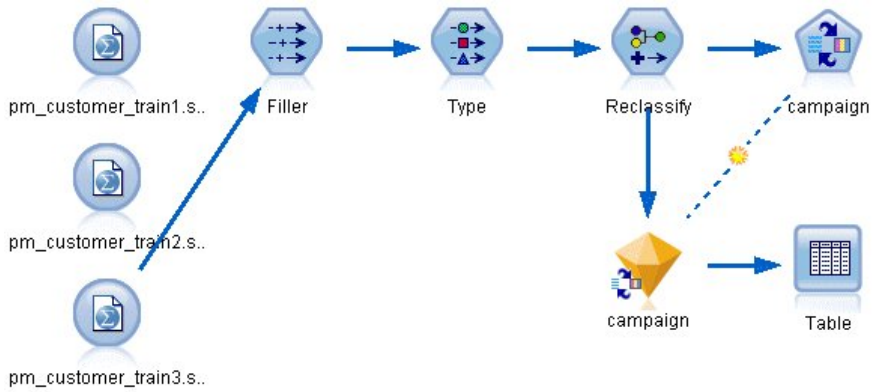


Figura 227. Anexando terceira origem de dados ao fluxo SLRM

9. Clique em **Executar** para re-criar o nugget modelo mais uma vez. Para visualizar seus detalhes, dê um duplo clique no nugget sobre a tela.
10. A guia Modelo agora mostra a precisão final estimada das predições para cada oferta.

    Como é possível ver, a precisão média caiu ligeiramente (de 86.9% para 85.4%) conforme você incluiu as origens de dados adicionais; no entanto, essa flutuação é uma quantia mínima e pode ser atribuída a pequenas anomalias dentro dos dados disponíveis.

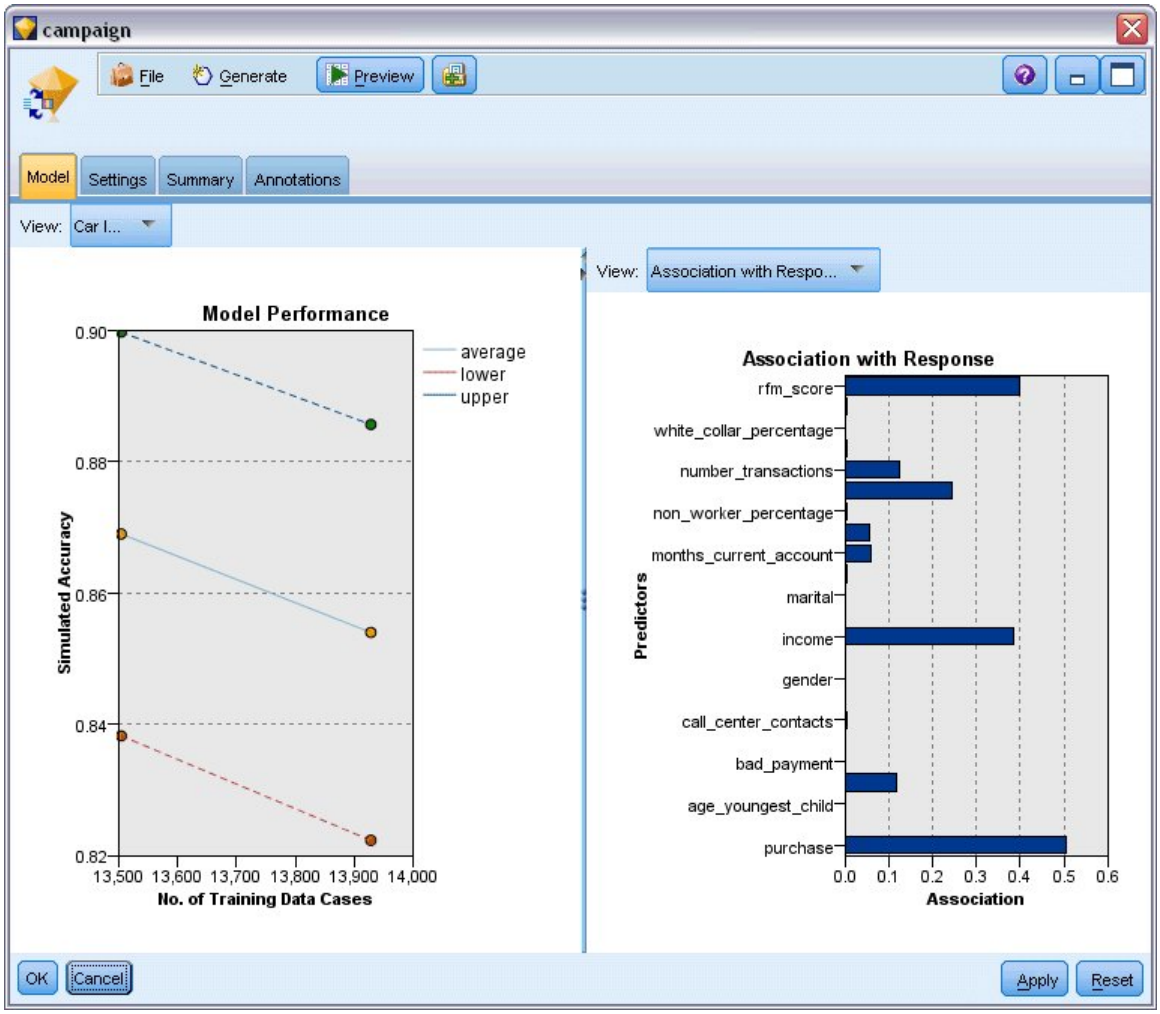
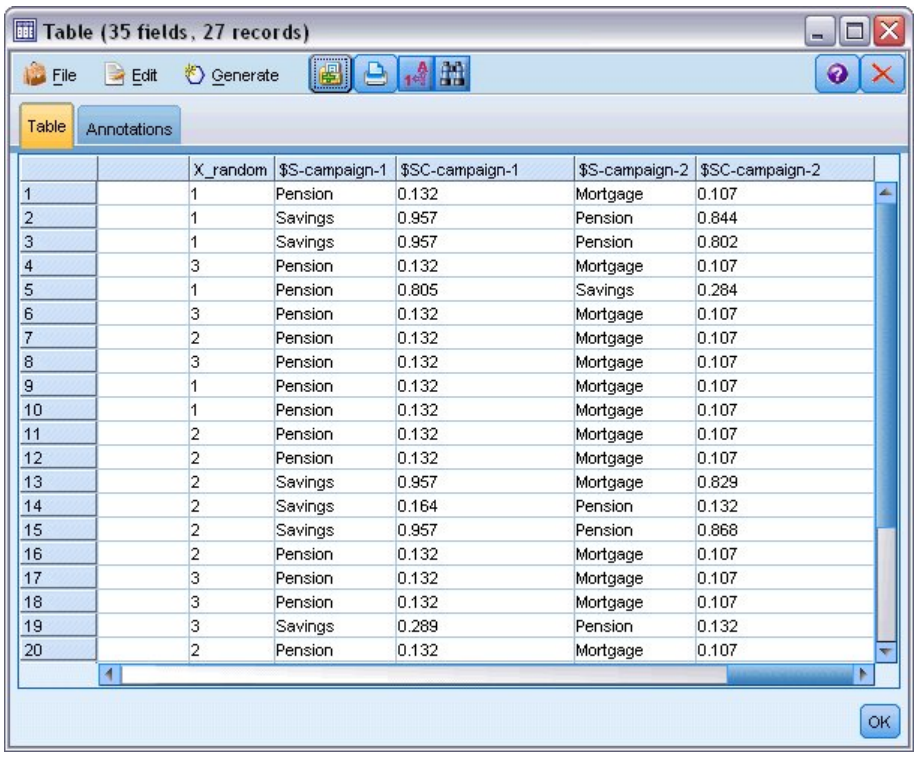


Figura 228. Nugget de modelo SLRM atualizado

11. Conecte um nó da Tabela ao último (terceiro) modelo gerado e execute o nó da Tabela.
12. Role em direção à direita da mesa. O show de previsões que oferece um cliente tem maior probabilidade de aceitar e a confiança que eles aceitarão, dependendo dos detalhes de cada cliente.

Por exemplo, na primeira linha da tabela mostrada, há apenas uma classificação de confiança de 13.2% (denotada pelo valor 0.132 na coluna *$SC-campaign-1*) que um cliente que tenha anteriormente contraído um empréstimo de carro aceitará uma pensão, se oferecida. No entanto, a segunda e a terceira linhas mostram mais dois clientes que também contraíram um empréstimo para automóveis; em seus casos, há uma confiança de 95,7% de que eles, e outros clientes com históricos semelhantes, abririam uma conta poupança se oferecida, e mais de 80% de confiança de que eles aceitariam uma pensão.



		X_random	$S-campaign-1	$SC-campaign-1	$S-campaign-2	$SC-campaign-2
1		1	Pension	0.132	Mortgage	0.107
2		1	Savings	0.957	Pension	0.844
3		1	Savings	0.957	Pension	0.802
4		3	Pension	0.132	Mortgage	0.107
5		1	Pension	0.805	Savings	0.284
6		3	Pension	0.132	Mortgage	0.107
7		2	Pension	0.132	Mortgage	0.107
8		3	Pension	0.132	Mortgage	0.107
9		1	Pension	0.132	Mortgage	0.107
10		1	Pension	0.132	Mortgage	0.107
11		2	Pension	0.132	Mortgage	0.107
12		2	Pension	0.132	Mortgage	0.107
13		2	Savings	0.957	Mortgage	0.829
14		2	Savings	0.164	Pension	0.132
15		2	Savings	0.957	Pension	0.868
16		2	Pension	0.132	Mortgage	0.107
17		3	Pension	0.132	Mortgage	0.107
18		3	Pension	0.132	Mortgage	0.107
19		3	Savings	0.289	Pension	0.132
20		2	Pension	0.132	Mortgage	0.107

Figura 229. Saída do modelo - ofertas e confidências previstas

Explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no *IBM SPSS Modelador Algorithms Guide*, disponível como arquivo PDF como parte do seu download do produto.

Observe também que esses resultados são baseados apenas nos dados de treinamento. Para avaliar o quão bem o modelo generaliza para outros dados no mundo real, você usaria um nó de partição para conter um subconjunto de registros para fins de teste e validação.

Capítulo 17. Prevendo Padrões De Empréstimo (Rede Bayesiana)

As redes bayesianas permitem construir um modelo de probabilidade combinando evidências observadas e registradas com conhecimentos do mundo real do "senso comum" para estabelecer a probabilidade de ocorrências usando atributos aparentemente desvinculados.

Este exemplo usa o fluxo denominado *bayes_bankloan.str*, que faz referência ao arquivo de dados denominado *bankloan.sav* Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador e podem ser acessados a partir do grupo de programas IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *bayes_bankloan.str* está no diretório *streams* .

Por exemplo, suponha que um banco esteja preocupado com o potencial de empréstimos para não serem repassados. Se os dados inadimplentes de empréstimos anteriores podem ser usados para prever quais potenciais clientes são passíveis de ter problemas para pagar empréstimos, esses clientes de "risco ruim" podem ser recusados um empréstimo ou oferecidos produtos alternativos.

Este exemplo foca em usar dados padrão de empréstimos existentes para prever potenciais padrões futuros, e olha para três tipos diferentes de modelo de rede Bayesiana para estabelecer qual é melhor em prever nesta situação.

Construindo o Fluxo

1. Inclua um nó de origem do Arquivo de Estatísticas apontando para *bankloan.sav* na pasta *Demos* .

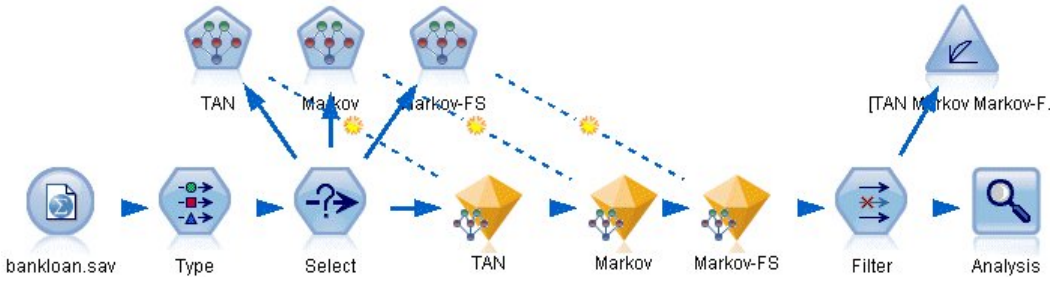


Figura 230. Fluxo de amostra da Rede Bayesiana

2. Inclua um nó Tipo no nó de origem e configure a função do campo **padrão** para **Destino**. Todos os outros campos devem ter seu papel configurado como **Entrada**.
3. Clique no botão **Ler Valores** para preencher a coluna *Valores* .

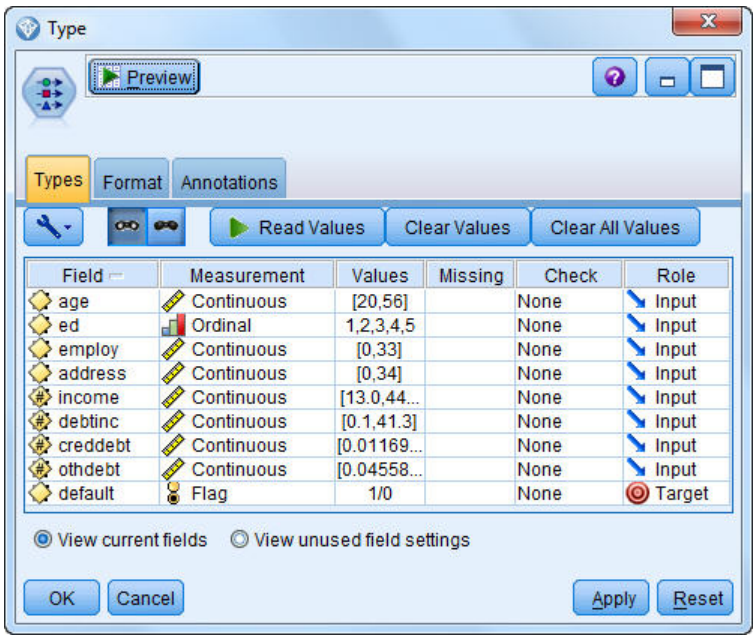


Figura 231. Selecionando o campo de destino

Os casos em que o alvo tem um valor nulo não são de uso ao construir o modelo. Você pode excluir esses casos para evitar que eles sejam usados na avaliação do modelo.

4. Inclua um nó Select no nó Type.
5. Para o Modo, selecione **Descarte**.
6. Na caixa de Condição, digite **default = '$null$'**.

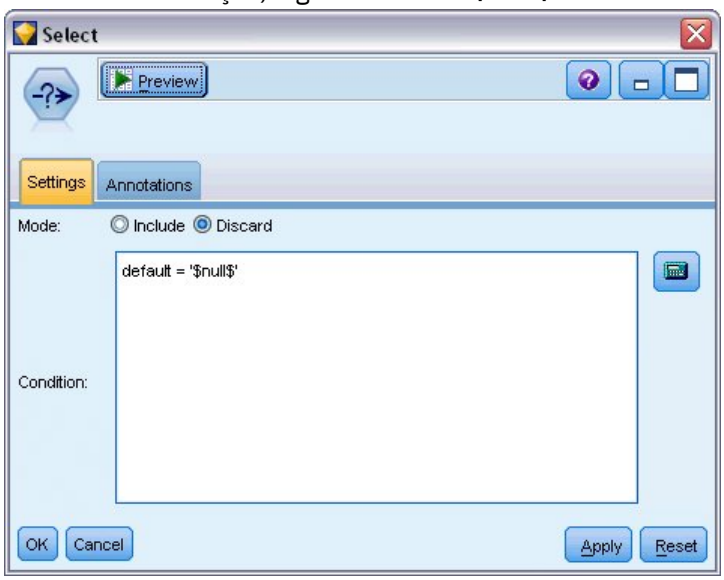


Figura 232. Descartando destinos nulos

Como você pode construir vários tipos diferentes de redes Bayesianas, vale a pena comparar várias para ver qual modelo fornece as melhores previsões. O primeiro a criar é um modelo Tree Augmented Naïve Bayes (TAN).

7. Conecte um nó da Rede Bayesiana ao nó Select.
8. Na guia Modelo, para o nome do Modelo, selecione **Customizado** e insira TAN na caixa de texto
9. Para Estrutura digite, selecione **TAN** e clique em **OK**.

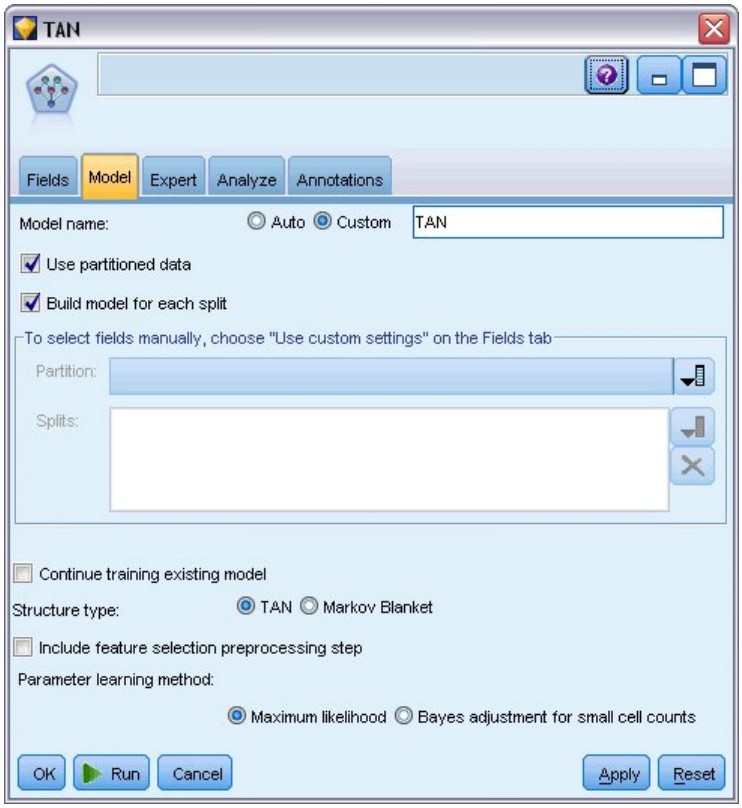


Figura 233. Criando um modelo Tree Augmented Naïve Bayes

O segundo tipo de modelo a construir tem uma estrutura de Markov Blanket.

10. Conecte um segundo nó da Rede Bayesiana ao nó Select.
11. Na guia Modelo, para o nome do Modelo, selecione **Customizado** e insira `Markov` na caixa de texto
12. Para Estrutura digite, selecione **Markov Blanket** e clique em **OK**.

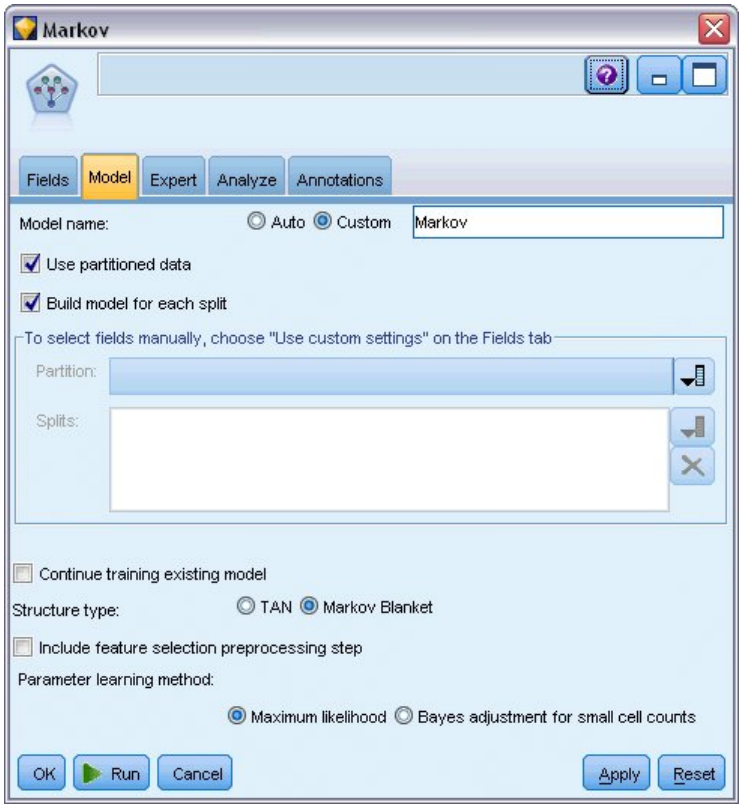


Figura 234. Criando um modelo Markov Blanket

O terceiro tipo de modelo a construir possui uma estrutura de Markov Blanket e também utiliza pré-processamento de seleção de recursos para selecionar as entradas que estão significativamente relacionadas à variável de destino.

13. Conecte um terceiro nó da Rede Bayesiana ao nó Select.
14. Na guia Modelo, para o nome do Modelo, selecione **Customizado** e insira `Markov-FS` na caixa de texto
15. Para Estrutura tipo, selecione **Markov Blanket**.
16. Selecione **Include o passo de pré-processamento da seleção de recursos** e clique em **OK**.

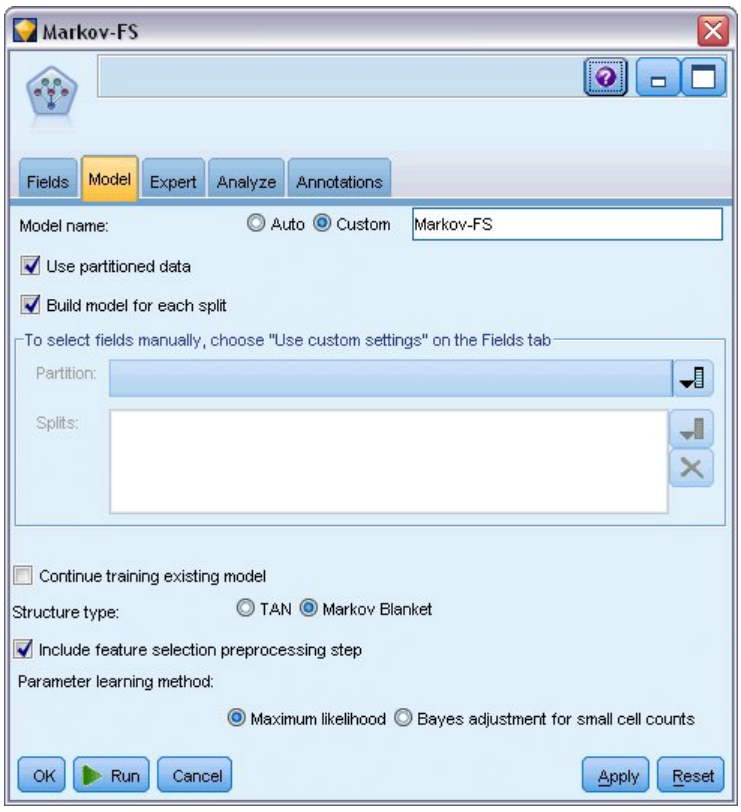


Figura 235. Criação de um modelo Markov Blanket com pré-processamento de Feature Selection

Navegando no modelo

1. Execute o fluxo para criar os nuggets do modelo, que são adicionados ao fluxo e à paleta de Modelos no canto superior direito. Para visualizar seus detalhes, dê um duplo clique sobre qualquer um dos nuggets do modelo no fluxo.

   A guia Modelo do nugget do modelo é dividida em duas áreas de janela. O painel esquerdo contém um gráfico de rede de nós que exibe a relação entre o alvo e seus preditores mais importantes, assim como a relação entre os preditores.

   O painel direito mostra a *Importância Preditadora*, que indica a importância relativa de cada preditor na estimativa do modelo, ou *Probabilidades Condicionais*, que contém o valor de probabilidade condicional para cada valor de nó e cada combinação de valores em seus nós pais.

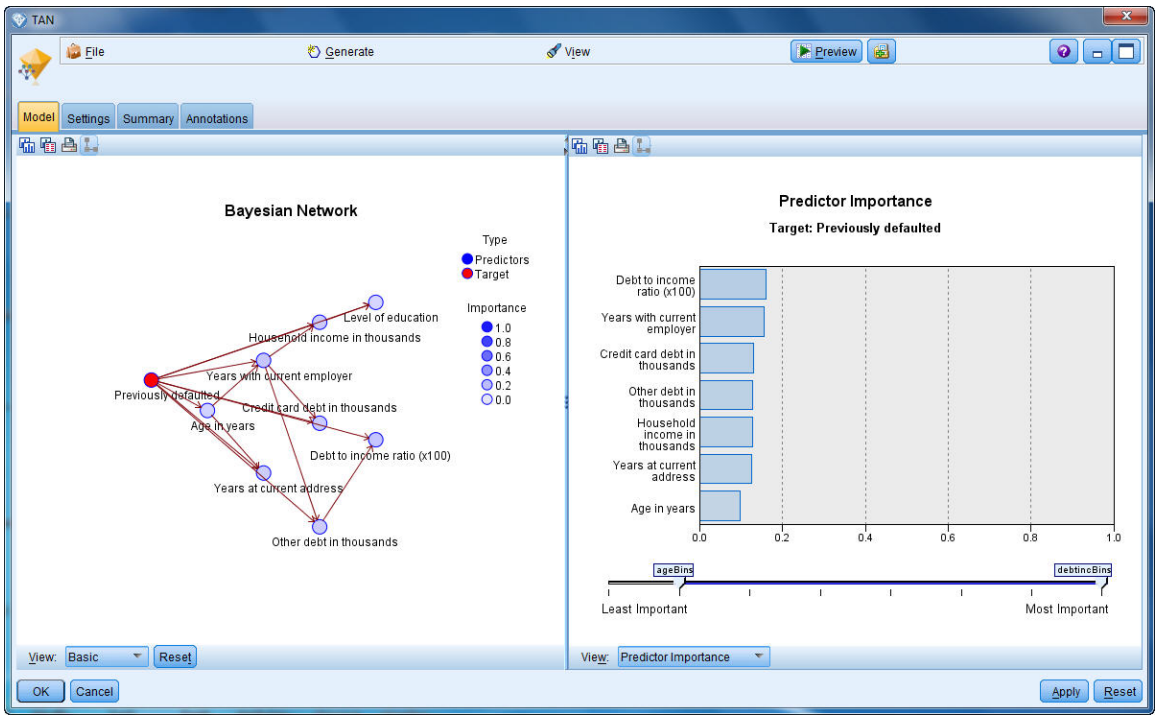


Figura 236. Como visualizar um modelo Tree Augmented Naïve Bayes

2. Conecte o nugget do modelo TAN ao nugget Markov (escolha **Substituir** no diálogo de aviso).
3. Conecte o nugget Markov ao nugget Markov-FS (escolha **Substituir** no diálogo de aviso).
4. Alinhe os três nuggets com o nó Select para facilitar a visualização.

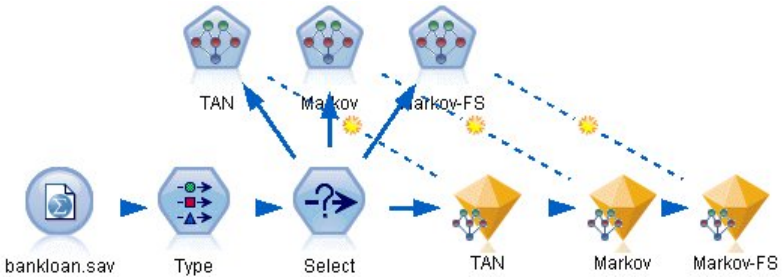


Figura 237. Alinhando os nuggets no riacho

5. Para renomear as saídas do modelo para obter clareza no gráfico de Avaliação que você estará criando, anexe um nó Filtro ao nugget modelo Markov-FS.
6. Na coluna direito *Campo* , renomear $B-default como TAN, $B1-default como Markov, e $B2-default como Markov-FS.

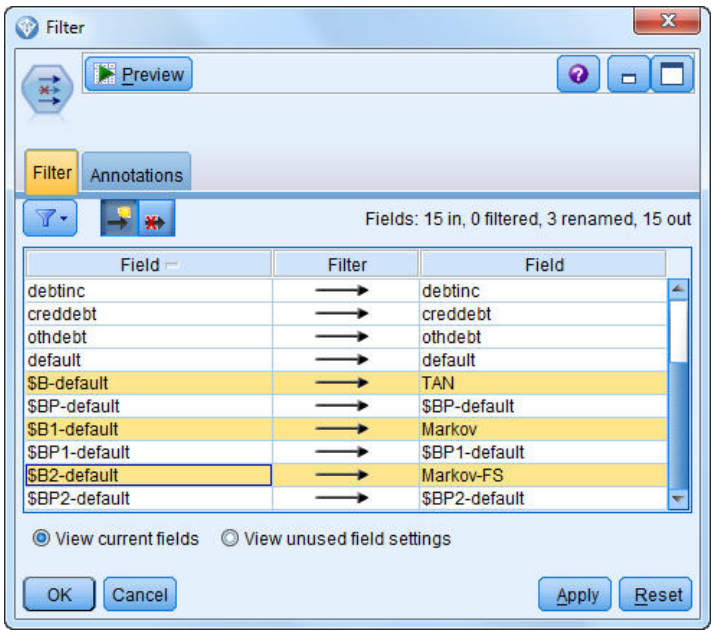


Figura 238. Nomes de campo do modelo Rename

Para comparar a precisão prevista dos modelos, é possível construir um gráfico de ganhos.

7. Conecte um nó gráfico de Avaliação ao nó Filtro e execute o nó do gráfico usando suas configurações padrão.

   O gráfico mostra que cada tipo de modelo produz resultados semelhantes; no entanto, o modelo Markov é um pouco melhor.

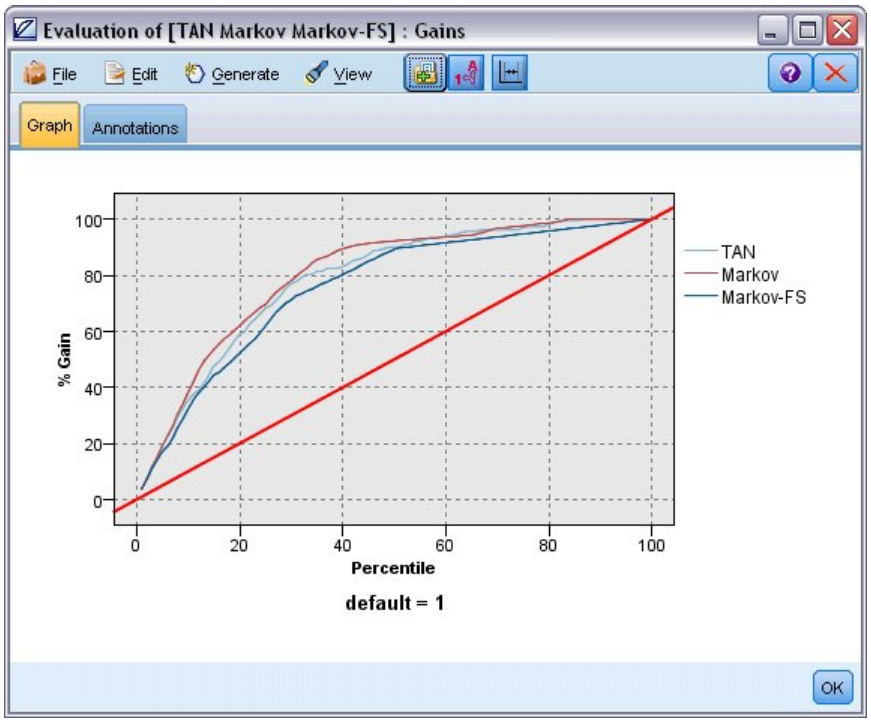


Figura 239. Avaliando precisão do modelo

Para verificar o quão bem cada modelo prevê, você poderia usar um nó de Análise em vez do gráfico de Avaliação. Isso mostra a precisão em termos de porcentagem para as previsões corretas e incorretas.

8. Anexar um nó de Análise ao nó Filtro e executar o nó de Análise usando suas configurações padrão.

Tal como acontece com o gráfico de Avaliação, isto mostra que o modelo de Markov é ligeiramente melhor em prever corretamente; no entanto, o modelo Markov-FS está apenas alguns pontos percentuais atrás do modelo Markov. Isso pode significar que seria melhor utilizar o modelo Markov-FS já que ele usa menos entradas para calcular seus resultados, economizando assim na coleta de dados e no tempo de entrada e no tempo de processamento.

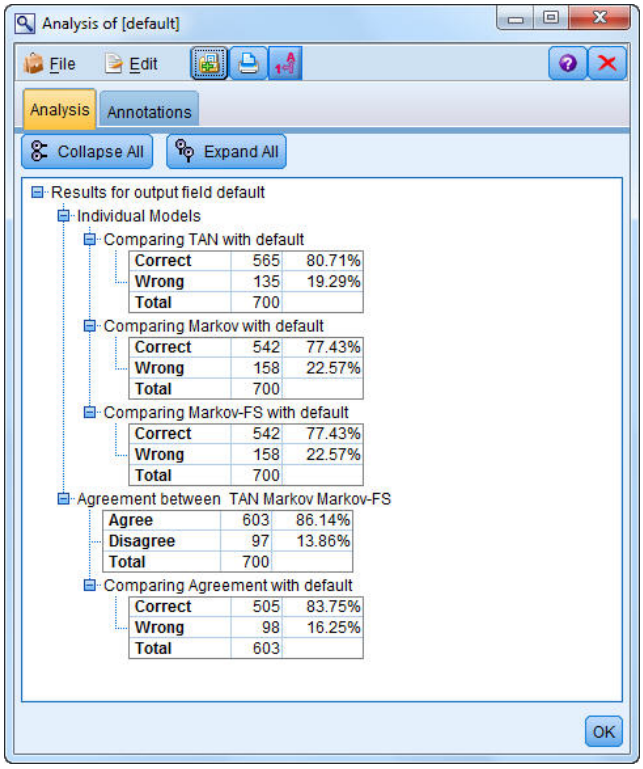


Figura 240. Analisando a precisão do modelo

Explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no *IBM SPSS Modelador Algorithms Guide*, disponível a partir do diretório *Documentação* do disco de instalação.

Observe também que esses resultados são baseados apenas nos dados de treinamento. Para avaliar o quão bem o modelo generaliza para outros dados no mundo real, você usaria um nó de partição para conter um subconjunto de registros para fins de teste e validação.

Capítulo 18. Retreinar um Modelo em uma Base Mensal (Rede Bayesiana)

As redes bayesianas permitem construir um modelo de probabilidade combinando evidências observadas e registradas com conhecimentos do mundo real do "senso comum" para estabelecer a probabilidade de ocorrências usando atributos aparentemente desvinculados.

Este exemplo usa o fluxo denominado *bayes_churn_retrain.str*, que faz referência aos arquivos de dados denominados *telco_Jan.sav* e *telco_Feb.sav* Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador e podem ser acessados a partir do grupo de programas IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *bayes_churn_retrain.str* está no diretório *streams* .

Por exemplo, suponhamos que um provedor de telecomunicações esteja preocupado com o número de clientes que está perdendo para os concorrentes (churn). Se os dados históricos do cliente podem ser usados para prever quais clientes são mais propensos a se agirem no futuro, esses clientes podem ser direcionados com incentivos ou outras ofertas para desestimulá-los de se transferir para outro provedor de serviços.

Este exemplo focaliza o uso de um dado de churn do mês existente para prever quais clientes podem provavelmente se agenar no futuro e, em seguida, adicionar os dados do mês seguinte para refinar e retreinar o modelo.

Construindo o Fluxo

1. Inclua um nó de origem do Arquivo de Estatísticas apontando para *telco_Jan.sav* na pasta *Demos* .

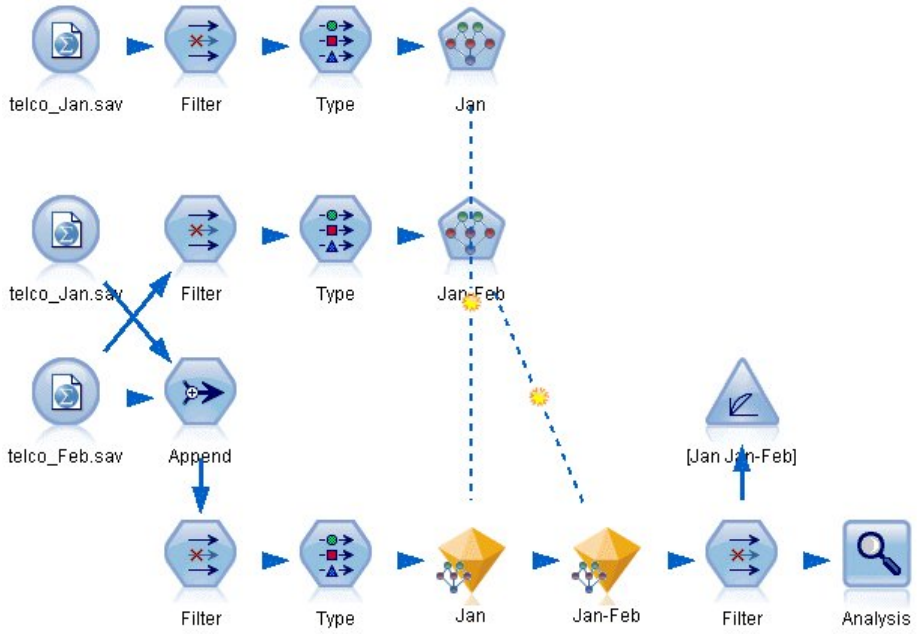


Figura 241. Fluxo de amostra da Rede Bayesiana

A análise anterior mostrou que vários campos de dados são de pouca importância ao prever churn. Esses campos podem ser filtrados a partir de seu conjunto de dados para aumentar a velocidade de processamento quando você está construindo e pontuando modelos.

2. Inclua um nó Filtro no nó Fonte.
3. Excluir todos os campos, exceto o *endereço*, *idade*, *churn*, *custcat*, *ed*, *empregar*, *gênero*, *marital*, *residir*, *aposentar*, e *tenure*.

4. Clique em **OK**.

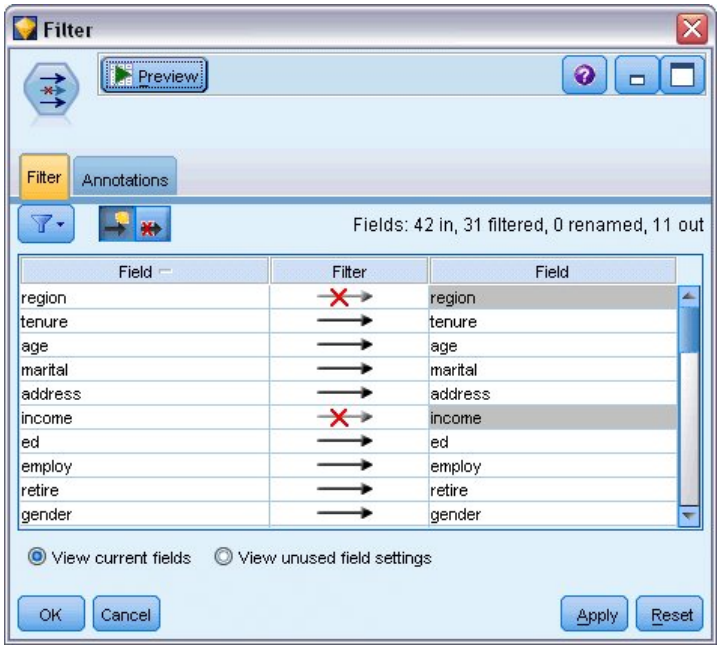


Figura 242. Filtrando campos desnecessários

5. Inclua um nó Tipo no nó Filtro.
6. Abra o nó Type e clique no botão **Ler Valores** para preencher a coluna *Valores* .
7. A fim de que o nó da Avaliação possa avaliar qual valor é verdadeiro e qual é falso, configure o nível de medição para o campo *churn* para **Flag**, e configure sua função para **Target**. Clique em **OK**.

Figura 243. Selecionando o campo de destino

Você pode construir vários tipos diferentes de redes bayesianas; no entanto, para este exemplo você vai construir um modelo de Tree Augmented Naïve Bayes (TAN). Isso cria uma grande rede e garante que você incluiu todos os links possíveis entre variáveis de dados, construindo assim um modelo inicial robusto.

8. Conecte um nó da Rede Bayesiana ao nó do Tipo.
9. Na guia Modelo, para o nome do Modelo, selecione **Customizado** e insira Jan na caixa de texto

10. Para Método de aprendizagem de parâmetro, selecione **Ajuste de Bayes para pequenas contagens de células**.
11. Clique em **Executar** . O nugget modelo é adicionado ao fluxo, e também à paleta de Models no canto superior direito.

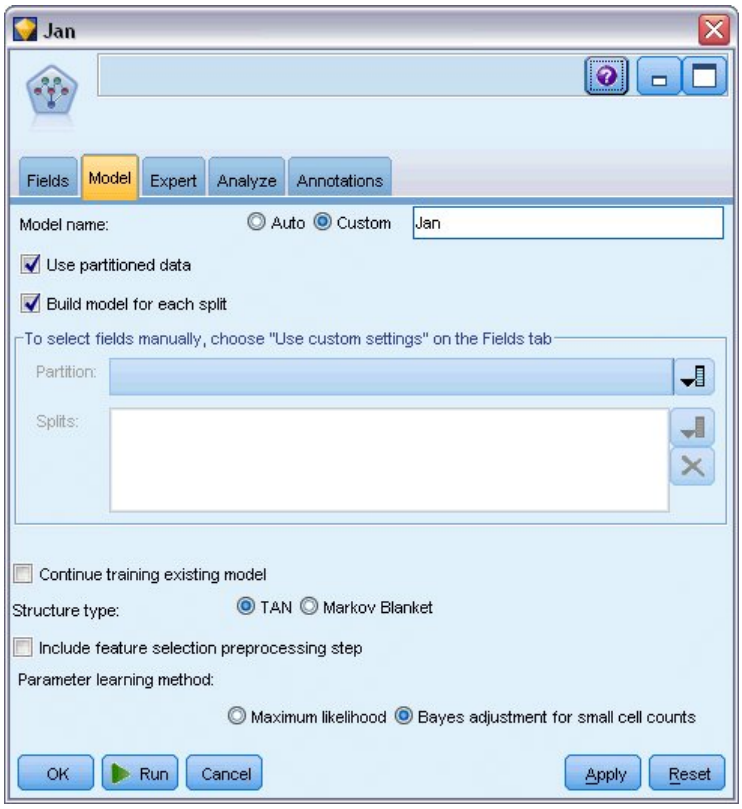


Figura 244. Criando um modelo Tree Augmented Naïve Bayes

12. Inclua um nó de origem do Arquivo de Estatísticas apontando para *telco_Feb.sav* na pasta *Demos* .
13. Conecte este novo nó de origem ao nó Filtro (no diálogo de aviso, escolha **Substituir** para substituir a conexão para o nó de origem anterior).

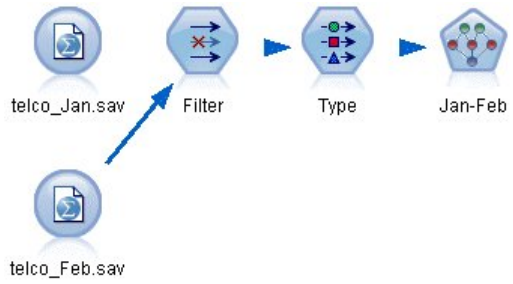


Figura 245. Como adicionar os dados do segundo mês

14. Na guia Modelo do nó Rede Bayesiana, para Nome do modelo, selecione **Customizado** e insira Jan-Feb na caixa de texto.
15. Selecione **Continuar treinando modelo existente**.
16. Clique em **Executar** . O modelo nugget sobrescreve o existente no fluxo, mas também é adicionado à paleta de Models no canto superior direito.

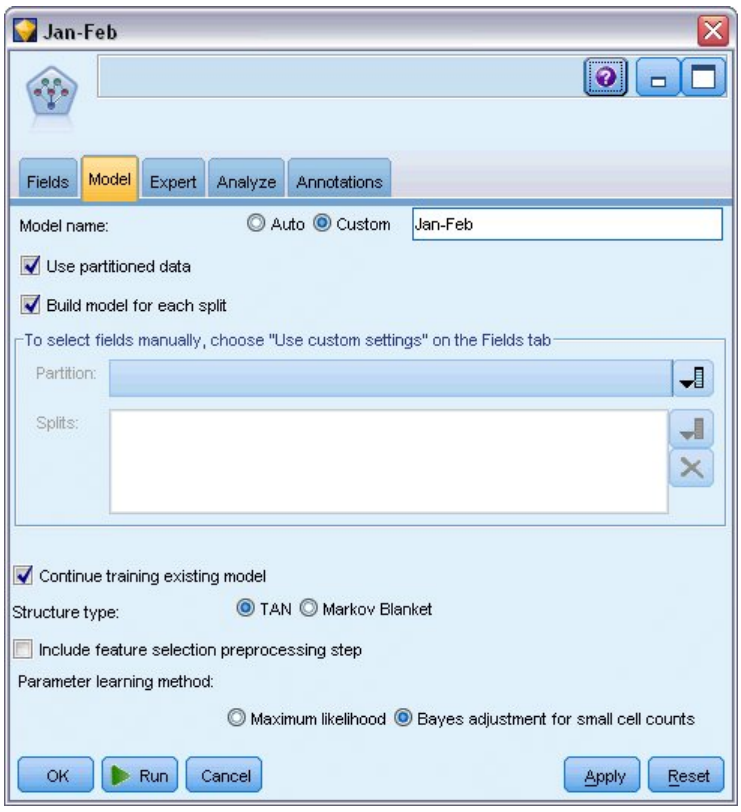


Figura 246. Retreinamento do modelo

Avaliando o Modelo

Para comparar os modelos, deve-se combinar os dois datasets.

1. Inclua um nó Anexo e anexe ambos os nós de origem *telco_Jan.sav* e *telco_Feb.sav* a ele.

Figura 247. Anexar as duas fontes de dados

2. Copie os nós Filtro e Type de anteriores no fluxo e cole-os sobre a tela do fluxo.
3. Conecte o nó Append ao nó do Filtro recentemente copiado.

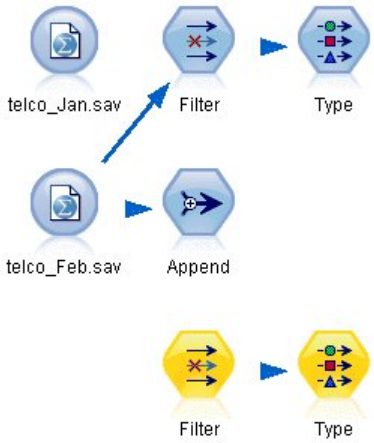


Figura 248. Colando os nós copiados no fluxo

Os nuggets para os dois modelos da Rede Bayesiana estão localizados na paleta de Models no canto superior direito.

4. Clique duas vezes no nugget do modelo Jan para trazê-lo para dentro do fluxo, e anexe-o ao nó do Tipo recém-copiado.
5. Conecte o nugget de modelo Jan-Feb já no fluxo para o nugget modelo de Jan.
6. Abra o nugget modelo de Jan.

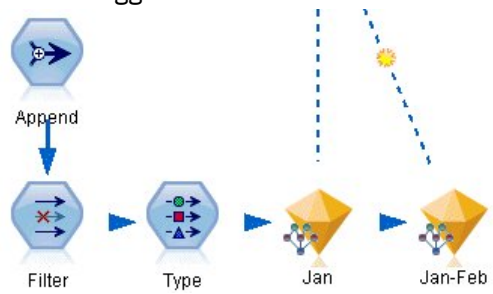


Figura 249. Adicionando os nuggets ao fluxo

A guia Nugget Modelo de Rede Bayesiana é dividida em duas colunas. A coluna esquerda contém um gráfico de rede de nós que exibe a relação entre o alvo e seus preditores mais importantes, assim como a relação entre os preditores.

A coluna da direita mostra a *Importância Preditadora*, que indica a importância relativa de cada preditor na estimativa do modelo, ou *Probabilidades Condicionais*, que contém o valor de probabilidade condicional para cada valor do nó e cada combinação de valores em seus nós pais.

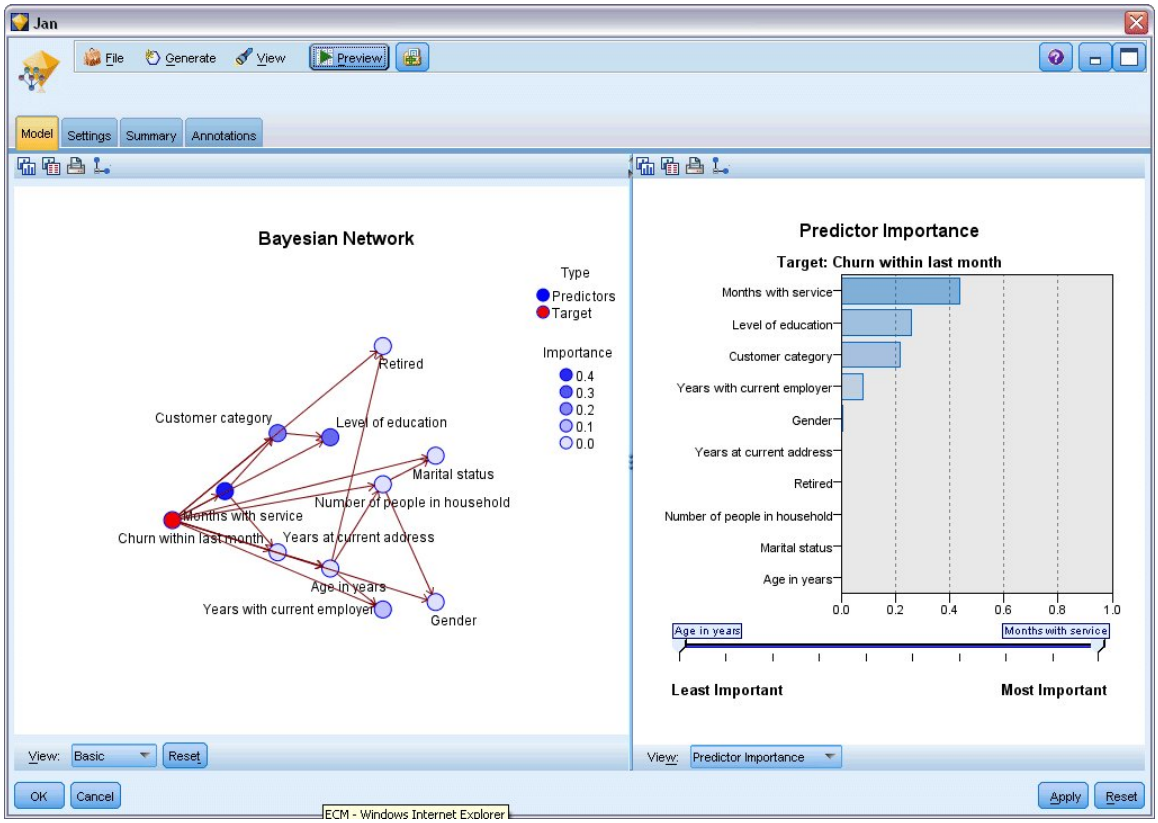


Figura 250. Modelo de Rede Bayesiana mostrando importância do preditor

Para exibir as probabilidades condicionais para qualquer nó, clique sobre o nó na coluna da esquerda. A coluna da direita é atualizada para mostrar os detalhes necessários.

As probabilidades condicionais são mostradas para cada bin que os valores de dados foram divididos em relativos aos nós pai e irmão do nó.

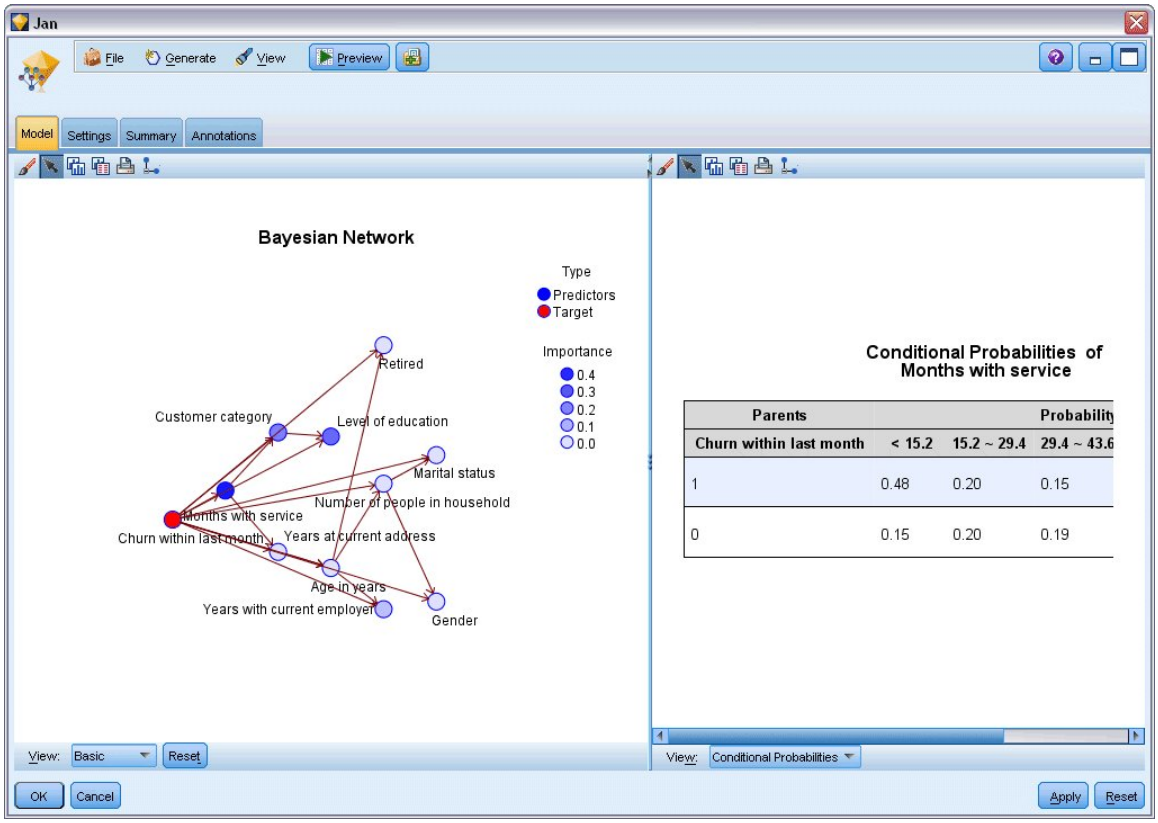


Figura 251. Modelo de Rede Bayesiana mostrando probabilidades condicionais

7. Para renomear as saídas de modelo para clareza, anexe um nó Filtro ao nugget modelo Jan-Feb.
8. Na coluna direito *Campo* , renomear $B-churn como Jan e $B1-churn como Jan-Feb.

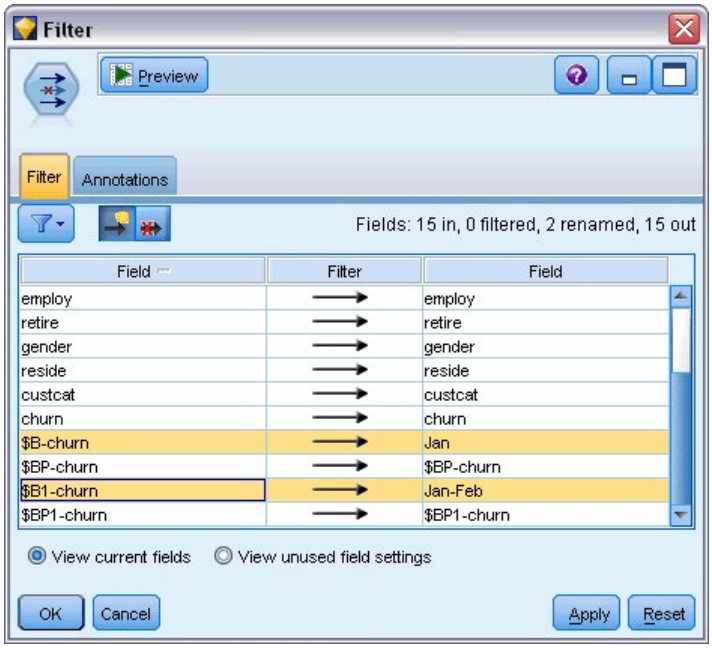


Figura 252. Nomes de campo do modelo Rename

Para verificar o quão bem cada modelo prevê churn, use um nó de Análise; isso mostra a precisão em termos de porcentagem para as previsões corretas e incorretas.

9. Anexar um nó de Análise ao nó Filtro.
10. Abra o nó da Análise e clique em **Executar**.

Isso mostra que ambos os modelos têm um grau semelhante de precisão ao prever churn.

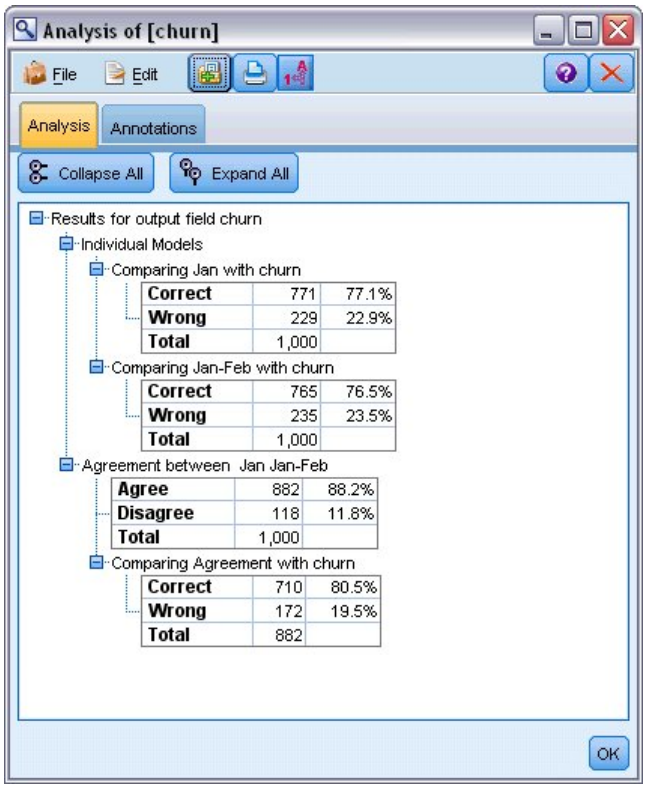


Figura 253. Analisando a precisão do modelo

Como uma alternativa ao nó de Análise, você pode usar um gráfico de Avaliação para comparar a precisão prevista dos modelos construindo um gráfico de ganhos.

11. Anexar um nó do gráfico de Avaliação ao nó Filtro.

e executar o nó do gráfico usando suas configurações padrão.

Assim como ocorre com o nó da Análise, o gráfico mostra que cada tipo de modelo produz resultados semelhantes; no entanto, o modelo reformado usando os dados de ambos os meses é um pouco melhor porque tem um nível de confiança maior em suas previsões.

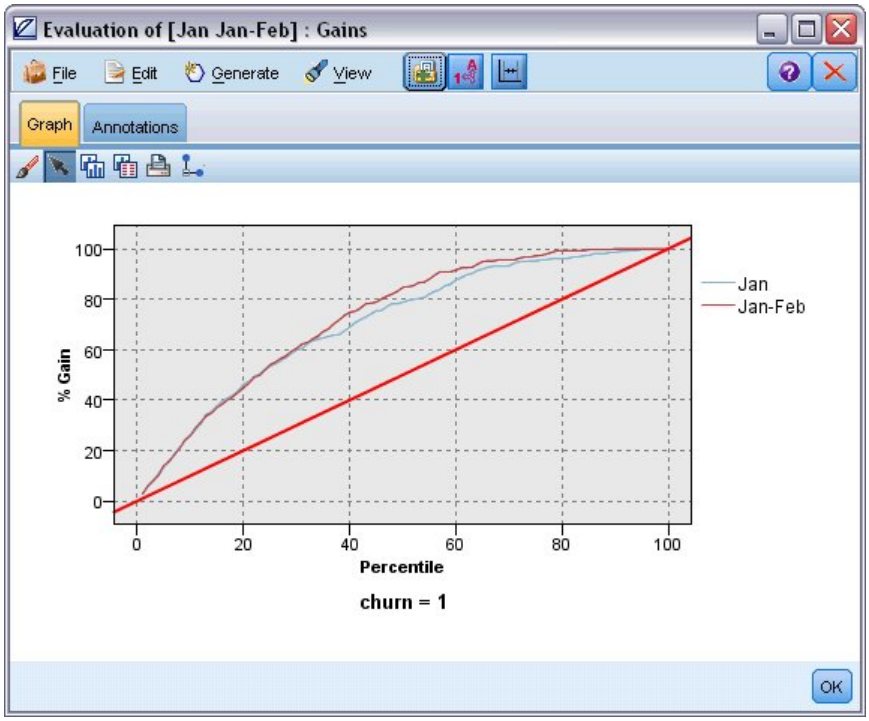


Figura 254. Avaliando precisão do modelo

Explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no *IBM SPSS Modelador Algorithms Guide*, disponível a partir do diretório *Documentação* do disco de instalação.

Observe também que esses resultados são baseados apenas nos dados de treinamento. Para avaliar o quão bem o modelo generaliza para outros dados no mundo real, você usaria um nó de partição para conter um subconjunto de registros para fins de teste e validação.

Capítulo 19. Promoção De Vendas No Varejo (Neural Net / C & RT)

Este exemplo trata de dados que descrevem as linhas de produtos de varejo e os efeitos da promoção nas vendas. (Este dado é fictício.) Seu objetivo neste exemplo é prever os efeitos de futuras promoções de vendas. Semelhante ao exemplo de monitoramento de condição, o processo de mineração de dados consiste na exploração, preparação de dados, treinamento e fases de teste.

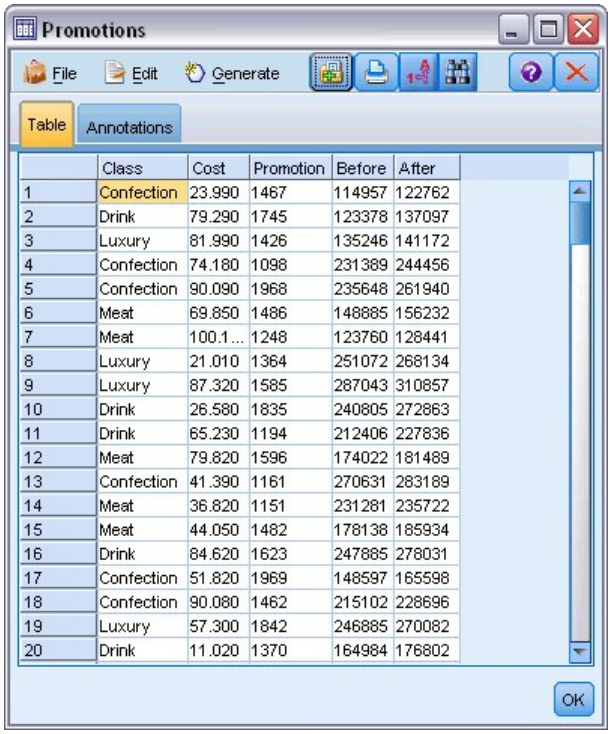
Este exemplo usa os fluxos denominados *goodsplot.str* e *goodslearn.str*, que fazem referência aos arquivos de dados denominados *GOODS1n* e *GOODS2n*. Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O fluxo *goodsplot.str* está na pasta *stream* , enquanto o arquivo *goodslearn.str* está no diretório *streams* .

Examinando os dados

Cada registro contém:

- *Classe*. Tipo de produto.
- *Custo*. Preço unitário.
- *Promoção*. Índice de valor gasto em uma promoção específica.
- *Antes*. Receita antes da promoção.
- *Depois*. Receita após a promoção.

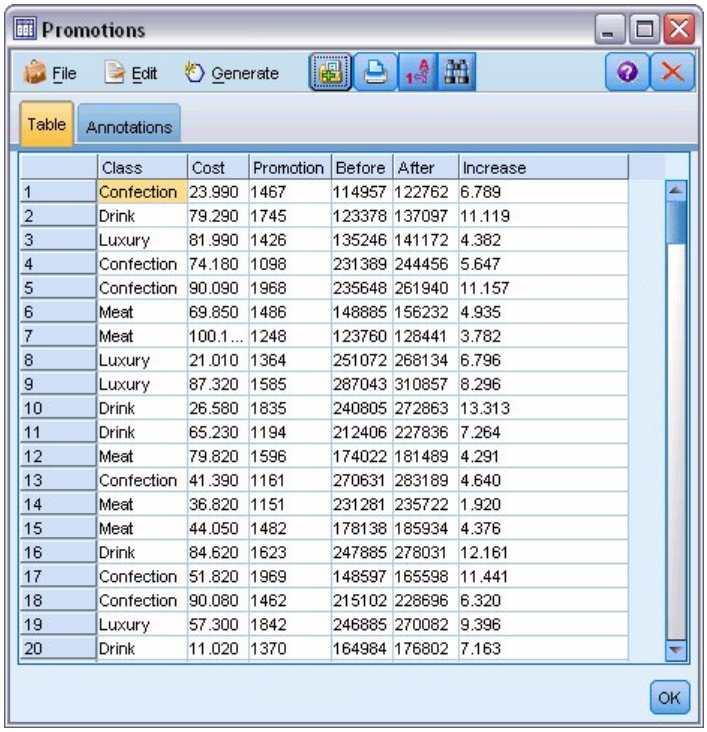
O fluxo *goodsplot.str* contém um fluxo simples para exibir os dados em uma tabela.. Os dois campos de receita (*Antes* e *Depois*) são expressos em termos absolutos; no entanto, parece provável que o aumento da receita após a promoção (e presumivelmente como resultado dela) seria uma figura mais útil.



	Class	Cost	Promotion	Before	After
1	Confection	23.990	1467	114957	122762
2	Drink	79.290	1745	123378	137097
3	Luxury	81.990	1426	135246	141172
4	Confection	74.180	1098	231389	244456
5	Confection	90.090	1968	235648	261940
6	Meat	69.850	1486	148885	156232
7	Meat	100.1...	1248	123760	128441
8	Luxury	21.010	1364	251072	268134
9	Luxury	87.320	1585	287043	310857
10	Drink	26.580	1835	240805	272863
11	Drink	65.230	1194	212406	227836
12	Meat	79.820	1596	174022	181489
13	Confection	41.390	1161	270631	283189
14	Meat	36.820	1151	231281	235722
15	Meat	44.050	1482	178138	185934
16	Drink	84.620	1623	247885	278031
17	Confection	51.820	1969	148597	165598
18	Confection	90.080	1462	215102	228696
19	Luxury	57.300	1842	246885	270082
20	Drink	11.020	1370	164984	176802

Figura 255. Efeitos da promoção nas vendas de produtos

goodsplot.str também contém um nó para derivar esse valor, expresso como uma porcentagem da renda antes da promoção, em um campo chamado *Increase* e exibe uma tabela mostrando esse campo..



	Class	Cost	Promotion	Before	After	Increase
1	Confection	23.990	1467	114957	122762	6.789
2	Drink	79.290	1745	123378	137097	11.119
3	Luxury	81.990	1426	135246	141172	4.382
4	Confection	74.180	1098	231389	244456	5.647
5	Confection	90.090	1968	235648	261940	11.157
6	Meat	69.850	1486	148885	156232	4.935
7	Meat	100.1...	1248	123760	128441	3.782
8	Luxury	21.010	1364	251072	268134	6.796
9	Luxury	87.320	1585	287043	310857	8.296
10	Drink	26.580	1835	240805	272863	13.313
11	Drink	65.230	1194	212406	227836	7.264
12	Meat	79.820	1596	174022	181489	4.291
13	Confection	41.390	1161	270631	283189	4.640
14	Meat	36.820	1151	231281	235722	1.920
15	Meat	44.050	1482	178138	185934	4.376
16	Drink	84.620	1623	247885	278031	12.161
17	Confection	51.820	1969	148597	165598	11.441
18	Confection	90.080	1462	215102	228696	6.320
19	Luxury	57.300	1842	246885	270082	9.396
20	Drink	11.020	1370	164984	176802	7.163

Figura 256. Aumento da receita após a promoção

Além disso, o fluxo exibe um histograma do aumento e um espalhador do aumento contra os custos de promoção despendidos, sobrepostos com a categoria de produto envolvido.

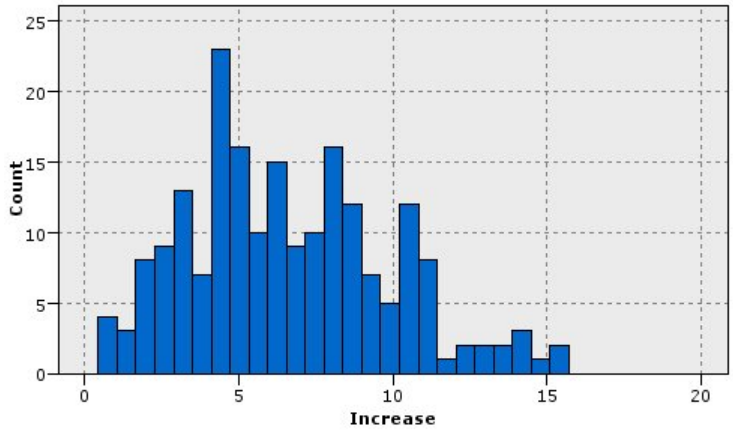


Figura 257. Histograma de aumento de receita

A dispersão mostra que, para cada classe de produto, existe uma relação quase linear entre o aumento da receita e o custo de promoção. Portanto, parece provável que uma árvore de decisão ou rede neural poderia prever, com razoável precisão, o aumento da receita dos outros campos disponíveis.

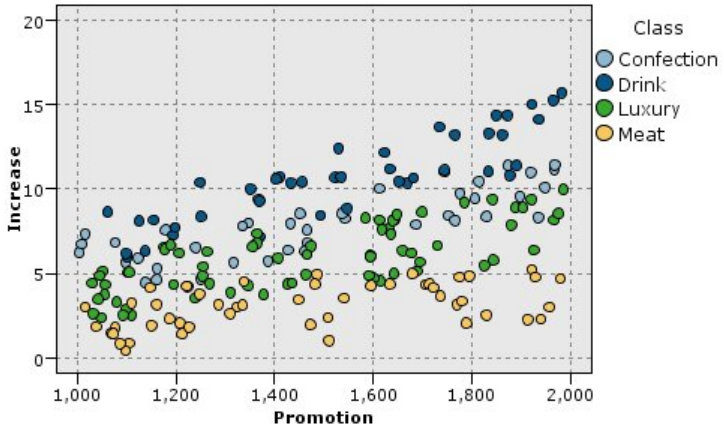


Figura 258. Aumento de receita versus despesas promocionais

Aprendizagem e teste

O fluxo `goodslearn.str` treina uma rede neural e uma árvore de decisão para fazer essa previsão de aumento de receita.

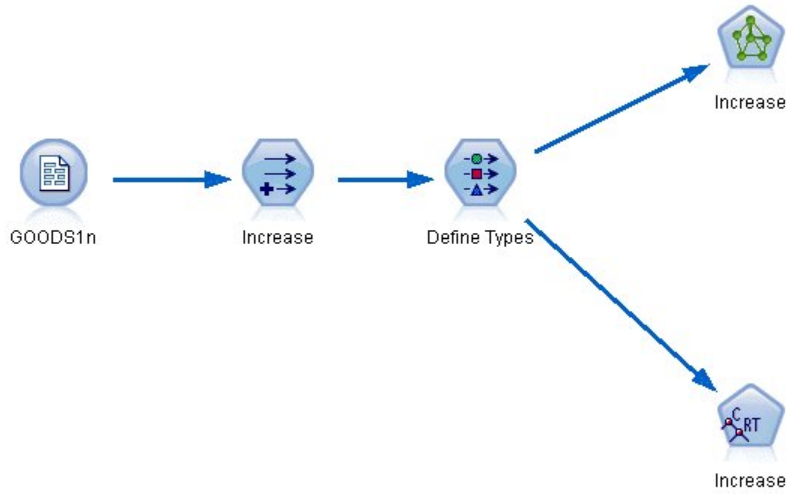


Figura 259. Modelando fluxo goodslearn.str

Uma vez executado os nós do modelo e gerado os modelos reais, é possível testar os resultados do processo de aprendizagem. Você faz isso conectando a árvore de decisão e rede em série entre o nó Tipo e um novo nó de Análise, alterando o arquivo de entrada (data) para G00DS2n, e executando o nó Análise. A partir da saída desse nó, em particular da correlação linear entre o aumento previsto e a resposta correta, você descobrirá que os sistemas treinados prevêem o aumento da receita com um alto grau de sucesso.

Uma maior exploração poderia concentrar-se nos casos em que os sistemas treinados cometem erros relativamente grandes; estes poderiam ser identificados traindo o aumento previsto de receita contra o aumento real. Os valores discrepantes neste gráfico podem ser selecionados usando os gráficos interativos dentro do ModeladorSPSS e, a partir de suas propriedades, pode ser possível ajustar a descrição dos dados ou o processo de aprendizagem para melhorar a precisão.

Capítulo 20. Monitoramento de Condição (Neural Net/ C5.0)

Este exemplo se refere ao monitoramento das informações de status de uma máquina e ao problema de reconhecer e prever estados de falha. Os dados são criados a partir de uma simulação fictícia e consistem em várias séries concatenadas medidas ao longo do tempo. Cada registro é um relatório instantâneo na máquina em termos do seguinte:

- *Horário*. Um número inteiro.
- *Energia*. Um número inteiro.
- *Temperatura*. Um número inteiro.
- *Pressão*. 0 se normal, 1 para um aviso de pressão momentânea.
- *Tempo de atividade*. Tempo desde a última manutenção.
- *Status*. Normalmente 0, alterações no código de erro no erro (101, 202 ou 303).
- *Resultado*. O código de erro que aparece nesta série temporal, ou 0 se não ocorrer nenhum erro. (Esses códigos estão disponíveis apenas em retrospectiva.)

Este exemplo usa os fluxos denominados *condplot.str* e *condlearn.str*, que fazem referência aos arquivos de dados denominados *COND1n* e *COND2n*. Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. Os arquivos *condplot.str* e *condlearn.str* estão no diretório de *streams*

Para cada série temporal, há uma série de registros a partir de um período de funcionamento normal seguido de um período que leva à falha, conforme mostrado na tabela a seguir:

Hora	Potência	Temperatura	Pressão	Tempo de atividade	Estado	Resultado
0	1059	259	0	404	0	0
1	1059	259	0	404	0	0
			...			
51	1059	259	0	404	0	0
52	1059	259	0	404	0	0
53	1007	259	0	404	0	303
54	998	259	0	404	0	303
			...			
89	839	259	0	404	0	303
105	834	259	0	404	303	303
0	965	251	0	209	0	0
1	965	251	0	209	0	0
			...			
51	965	251	0	209	0	0
52	965	251	0	209	0	0
53	938	251	0	209	0	104
54	936	251	0	209	0	104

Hora	Potência	Temperatura	Pressão	Tempo de atividade	Estado	Resultado
			...			
208	644	251	0	209	0	104
209	640	251	0	209	104	104

O processo a seguir é comum à maioria dos projetos de mineração de dados:

- Examine os dados para determinar quais atributos podem ser relevantes para a predição ou reconhecimento dos estados de interesse.
- Reter esses atributos (se já estiverem presentes) ou derivar e adicioná-los aos dados, se necessário.
- Use os dados resultantes para treinar regras e redes neurais.
- Teste os sistemas treinados usando dados de teste independentes.

Examinando os dados

O arquivo *condplot.str* ilustra a primeira parte do processo.. Ele contém um fluxo que trama uma série de gráficos. Se a série temporal de temperatura ou energia contiver padrões visíveis, você poderá diferenciar as condições de erro iminentes ou possivelmente prever sua ocorrência. Tanto para a temperatura como para a potência, o fluxo abaixo de tramas a série temporal associada aos três diferentes códigos de erro em gráficos separados, rendendo seis gráficos. Nós selecionados separam os dados associados aos diferentes códigos de erro.

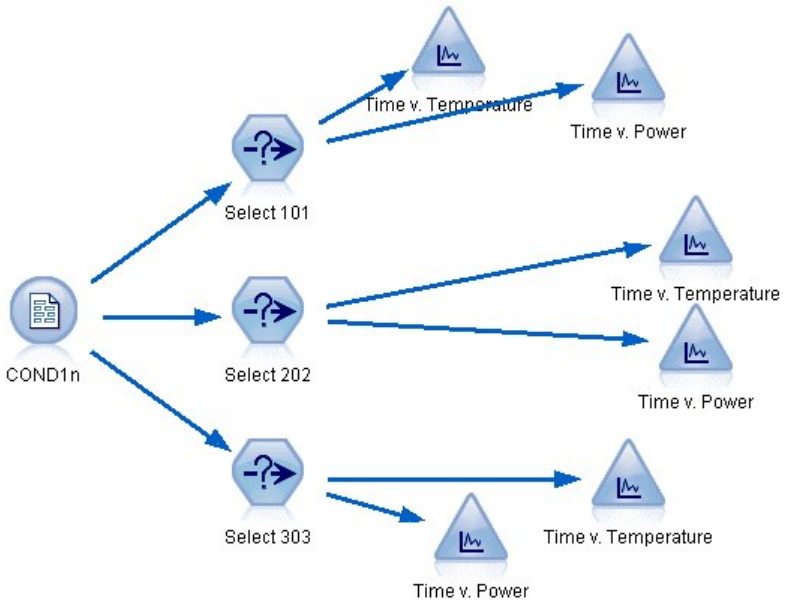


Figura 260. Fluxo de condplot

Os resultados deste fluxo são mostrados nesta figura.

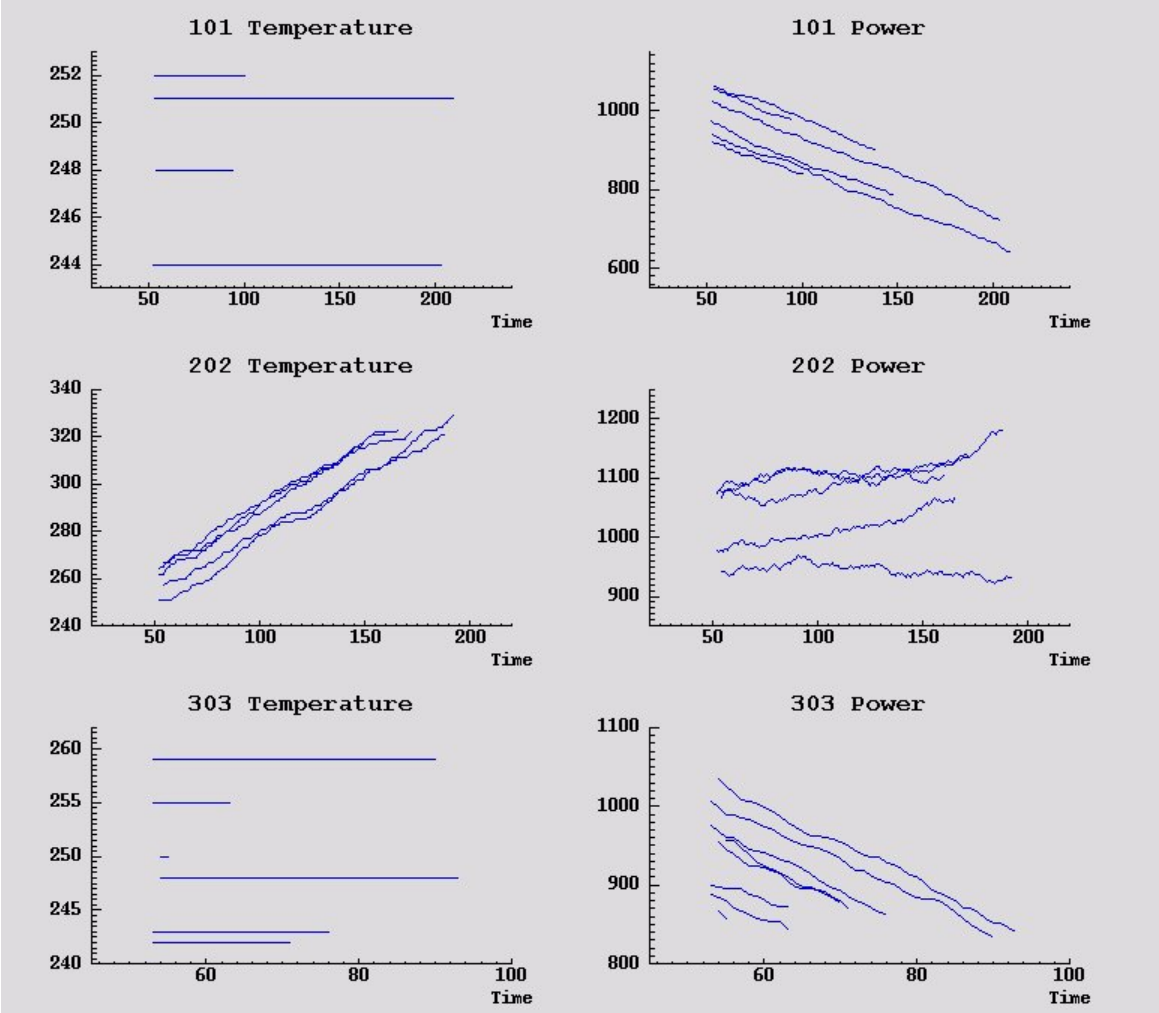


Figura 261. Temperatura e poder ao longo do tempo

Os gráficos exibem claramente padrões que distinguem erros 202 de erros 101 e 303. Os 202 erros mostram elevação da temperatura e oscilação do poder ao longo do tempo; os outros erros não. No entanto, os padrões que distinguem erros 101 de 303 são menos claros. Ambos os erros mostram uma temperatura uniforme e uma queda na potência, mas a queda na potência parece mais acentuada para erros 303.

Com base nesses gráficos, parece que a presença e a taxa de mudança para temperatura e energia, bem como a presença e o grau de flutuação, são relevantes para prever e distinguir falhas. Esses atributos devem, portanto, ser adicionados aos dados antes de aplicar os sistemas de aprendizagem.

Preparação de Dados

Com base nos resultados da exploração dos dados, o fluxo *condlearn.str* deriva os dados relevantes e aprende a prever falhas.

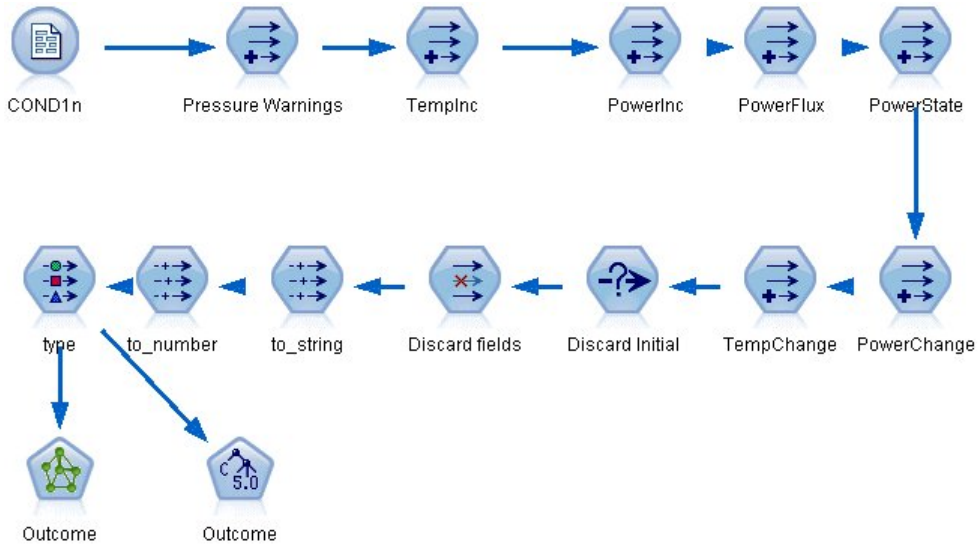


Figura 262. Fluxo de Aprenda

O fluxo usa vários nós de Derivação para preparar os dados para modelagem.

- **Nó de Arquivo Variável**. Lê o arquivo de dados *COND1n*.
- **Derivar Avisos de Pressão**. Conta o número de avisos de pressão momentâneos. Reconfigure quando o horário voltar para 0.
- **Derivar TempInc** Calcula a taxa momentânea de mudança de temperatura usando `@DIFF1`.
- **Derivar PowerInc**.. Calcula a taxa momentânea de mudança de potência usando `@DIFF1`.
- **Derivar PowerFlux**. Um sinalizador, true se a potência variou em direções opostas no último registro e neste; isto é, para um pico ou vale de energia.
- **Derivar PowerState**.. Um estado que começa como *Estável* e alterna para *Flutuando* quando dois fluxos de energia sucessivos são detectados. Alterna de volta para *Estável* somente quando não há fluxo de energia por cinco intervalos de tempo ou quando o *Horário* é reconfigurado.
- **PowerChange**. Média de *PowerInc* nos últimos cinco intervalos de tempo.
- **TempChange**. Média de *TempInc* nos últimos cinco intervalos de tempo.
- **Descartar Inicial (Selecionar)**. Descarta o primeiro registro de cada série temporal para evitar grandes saltos (incorretos) em *Energia* e *Temperatura* nos limites.
- **Campos de descarte**. Reduz os registros para *Tempo de atividade*, *Status*, *Resultado*, *Avisos de pressão*, *PowerState*, *PowerChange* e *TempChange*.
- **Tipo.** Define a função de *Outcome* como **Target** (o campo para prever). Além disso, define o nível de medida de *Resultado* como **Nominal**, *Avisos de pressão* como **Contínuo**e *PowerState* como **Sinalizador**.

Aprendizado

Executar o fluxo em *condlearn.str* treina a regra C5.0 e a rede neural (rede). A rede pode levar algum tempo para treinar, mas o treinamento pode ser interrompido cedo para salvar uma rede que produza resultados razoáveis. Uma vez que o aprendizado esteja concluído, a aba Modelos na parte superior direita dos gerentes janela pisca para alertá-lo de que duas novas nuggets foram criadas: uma representa a rede neural e uma representa a regra.

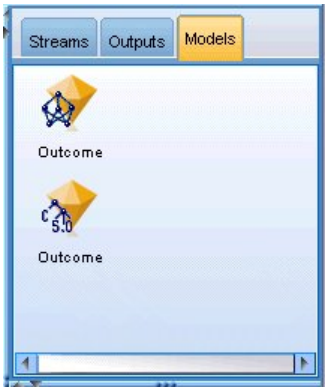


Figura 263. Gerente de modelos com nuggets modelo

Os nuggets do modelo também são adicionados ao fluxo existente, possibilitando-nos testar o sistema ou exportar os resultados do modelo. Neste exemplo, testaremos os resultados do modelo.

Testando

As nuggets do modelo são adicionadas ao fluxo, ambas conectadas ao nó Type.

1. Reposicionar os nuggets como mostrado, de modo que o nó Type se conecte ao nugget net neural, que se conecta ao nugget C5.0 .
2. Anexe um nó de análise ao nugget C5.0.
3. Edite o nó de origem original para ler o arquivo *COND2n* (em vez de *COND1n*), como *COND2n* contém dados de teste não vistos.

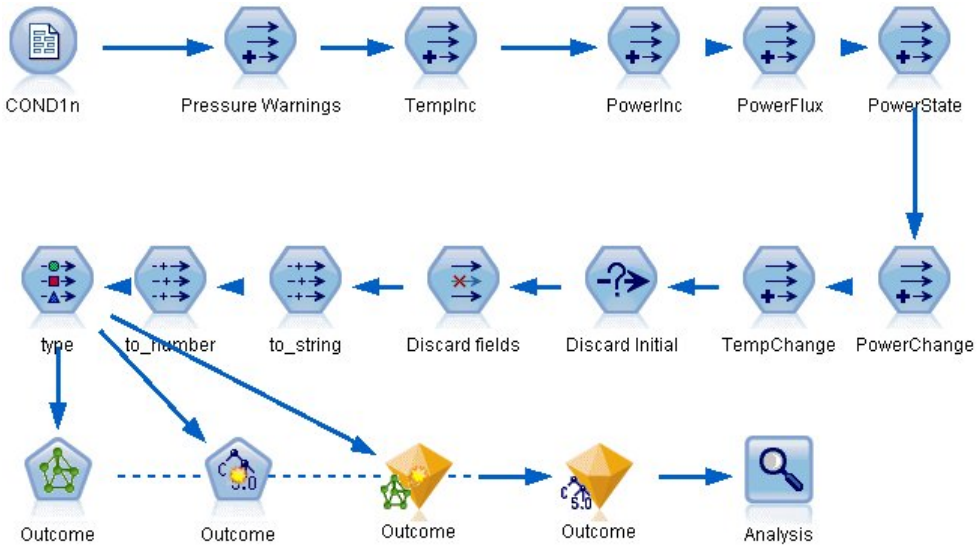


Figura 264. Testando a rede treinada

4. Abra o nó Análise e clique em Executar.

Isso gera números que refletem a precisão da rede treinada e da regra.

Capítulo 21. Classificando Os Clientes De Telecomunicações (Análise Discriminante)

A análise discriminante é uma técnica estatística para classificação de registros com base em valores de campos de entrada. Ela é semelhante a uma regressão linear, mas usa um campo de destino categórico ao invés de um campo numérico.

Por exemplo, suponha que um provedor de Telecomunicações tenha segmentado sua base de clientes por padrões de uso de serviço, categorizando os clientes em quatro grupos. Se os dados demográficos puderem ser usados para prever a associação ao grupo, será possível customizar ofertas para clientes em potencial individuais.

Este exemplo usa o fluxo denominado *telco_custcat_discriminant.str*, que faz referência ao arquivo de dados denominado *telco.sav* Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *telco_custcat_discriminant.str* está no diretório *streams* .

O exemplo se concentra no uso de dados demográficos para prever os padrões de uso. O campo de destino *custcat* possui quatro valores possíveis que correspondem aos quatro grupos de clientes, da seguinte forma:

Valor	**Rótulo**
1	Serviço Básico
2	E-Service
3	Serviço Plus
4	Serviço total

Criando o Fluxo

1. Primeiro, configure as propriedades do fluxo para mostrar etiquetas de variáveis e de valor na saída. Nos menus, escolha:

   Arquivo > **Propriedades do Fluxo ...** > **Opções** > **Geral**

2. Certise-se de que **Exibir etiquetas de campo e de valor na saída** esteja selecionado e clique em **OK**.

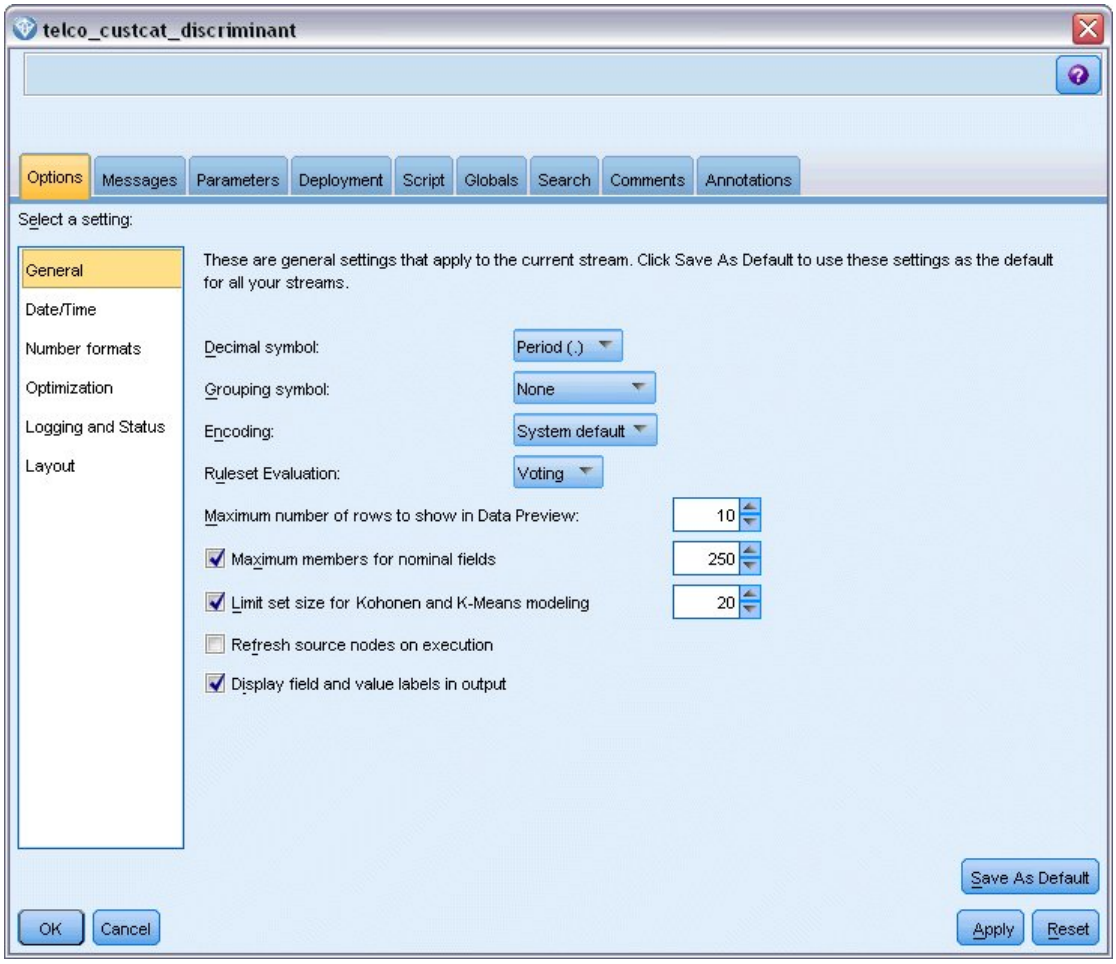


Figura 265. Propriedades do Fluxo

3. Inclua um nó de origem do Arquivo de Estatísticas apontando para *telco.sav* na pasta *Demos* .

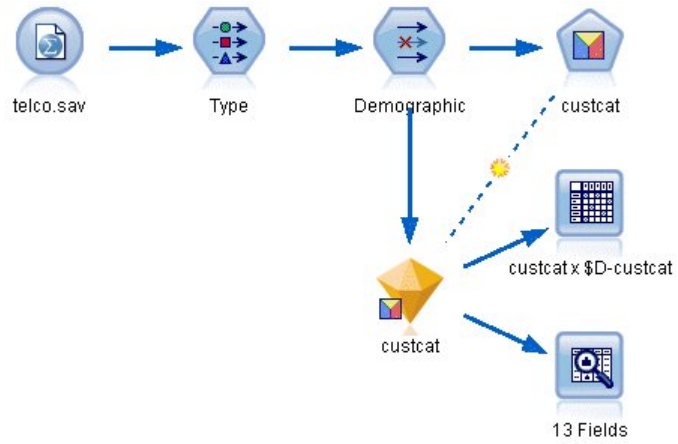


Figura 266. Fluxo de amostra para classificar os clientes usando análise discriminante

a. Adiciize um nó Tipo e clique em **Valores de leitura**, certificando-se de que todos os níveis de medição estão configurados corretamente. Por exemplo, a maioria dos campos com valores 0 e 1 pode ser considerada como sinalizadores.

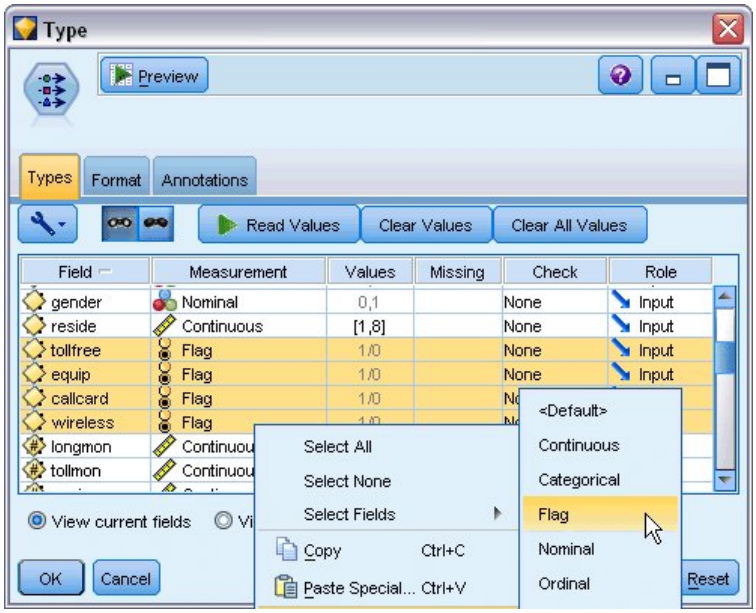


Figura 267. Configurando o nível de medição para vários campos

Sugestão: Para alterar propriedades para vários campos com valores semelhantes (como 0/ 1), clique no cabeçalho da coluna *Valores* para classificar campos por valor e, em seguida, mantenha a tecla shift ao utilizar as teclas do mouse ou seta para selecionar todos os campos que deseja alterar. Em seguida, é possível clicar com o botão direito do mouse na seleção para alterar o nível de medição ou outros atributos dos campos selecionados.

Observe que o *gênero* é considerado mais corretamente como um campo com um conjunto de dois valores, em vez de um sinalizador, portanto, deixe seu valor de medição como **Nominal**.

b. Configure a função para o campo *custcat* como **Destino**. Todos os outros campos devem ter seu papel configurado como **Entrada**.

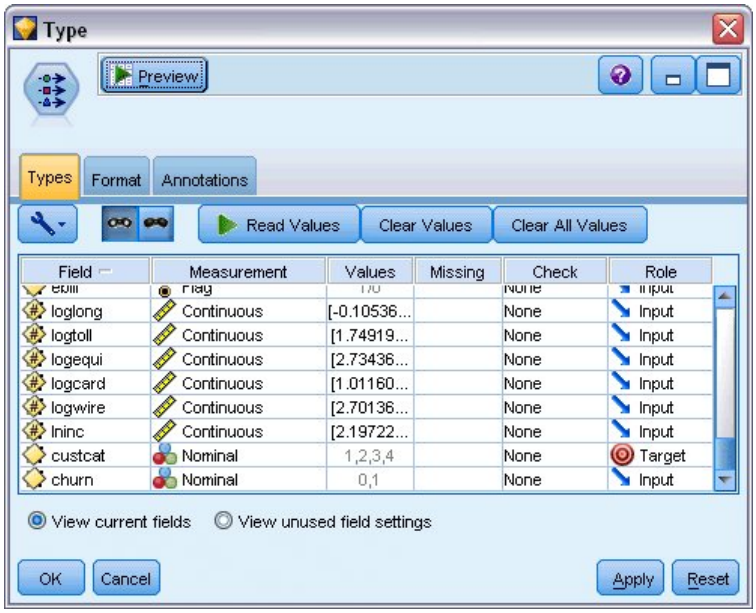


Figura 268. Configurando função de campo

Uma vez que este exemplo se concentra em demografia, use um nó Filtro para incluir apenas os campos relevantes (*região*, *idade*, *marital*, *endereço*, *renda*, *ed*, *empregar*, *aposentadoria*, *gênero*, *residir*, e *custcat*). Outros campos podem ser excluídos com a finalidade desta análise.

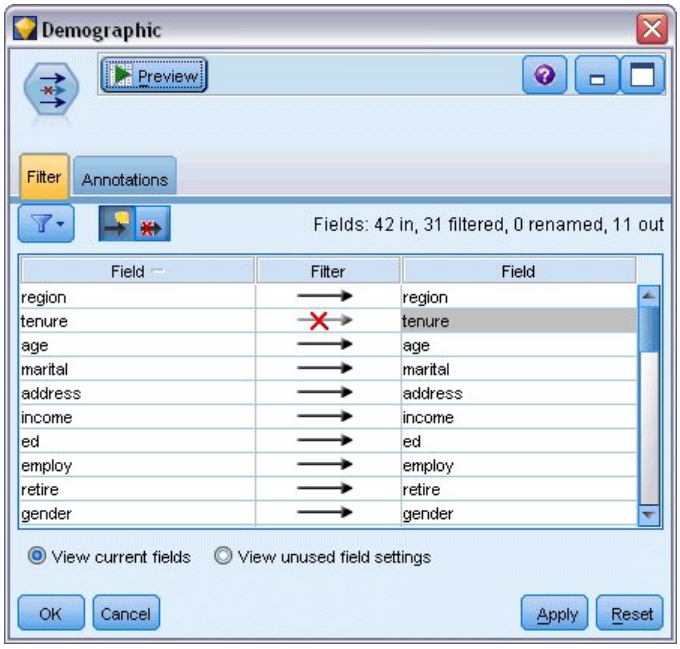


Figura 269. Filtragem em campos demográficos

(Alternativamente, você poderia alterar a função para **Nenhum** para esses campos em vez de excluí-los, ou selecionar os campos que deseja utilizar no nó de modelagem.)

4. No nó Discriminante, clique na guia Modelo e selecione o método **Stepwise** .

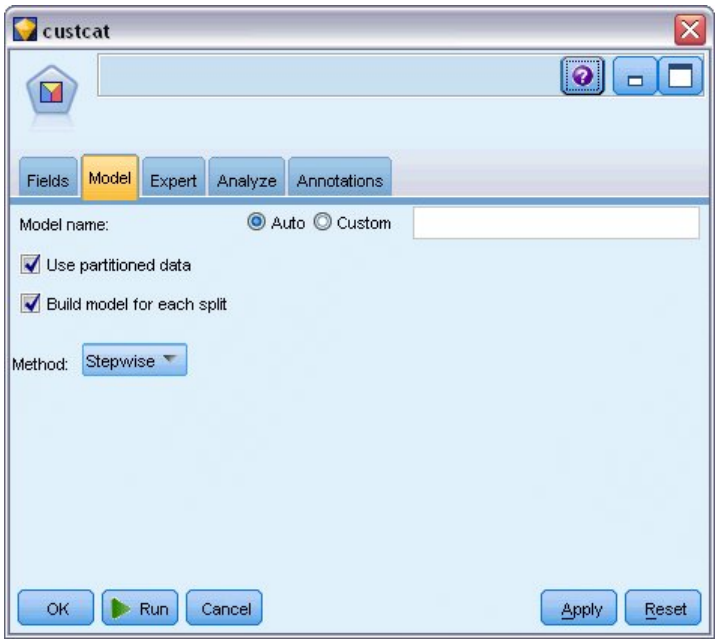


Figura 270. Escolhendo opções de modelo

5. Na guia Expert, configure o modo para **Expert** e clique em **Output**.
6. Selecione **Tabela de resumo**, **Mapa Territorial**e **Resumo de etapas** na caixa de diálogo de Saída Avançada, em seguida, clique em **OK**.

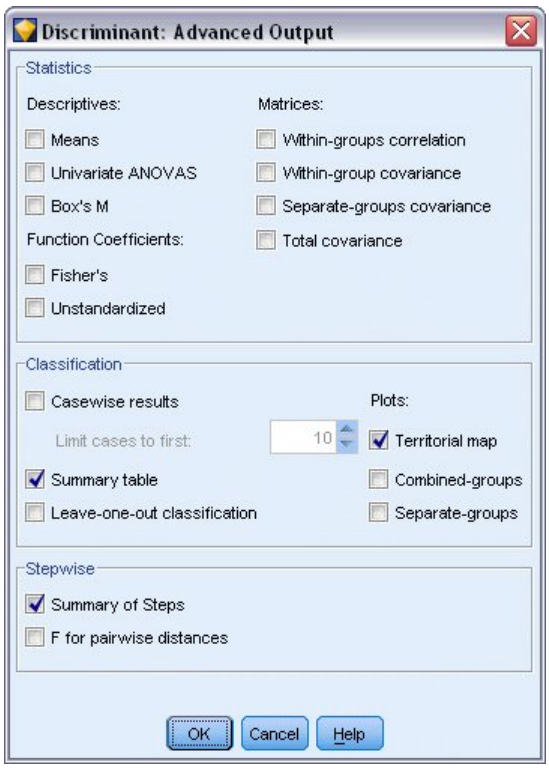


Figura 271. Escolhendo opções de saída

Examinando o modelo

1. Clique em **Executar** para criar o modelo, que é adicionado ao fluxo e à paleta de Models no canto superior direito. Para visualizar seus detalhes, dê um duplo clique sobre o nugget modelo no fluxo.

   A guia Resumo mostra (entre outras coisas) o alvo e a lista completa de entradas (campos do preditor) submetidos para consideração.

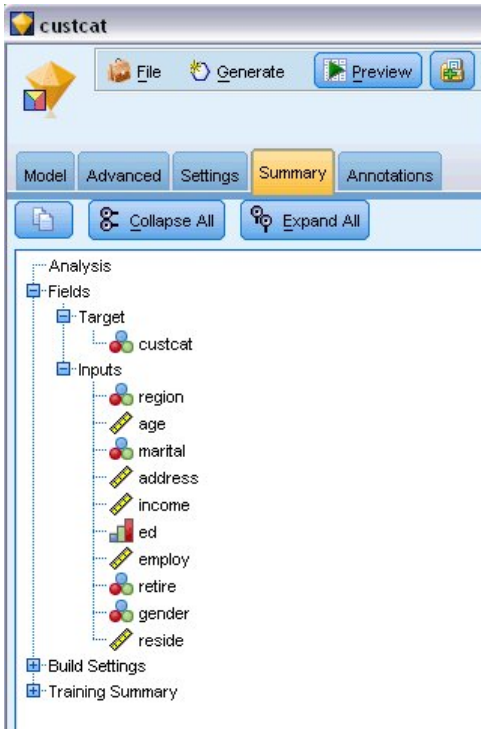


Figura 272. Sumário modelo mostrando campos de destino e entrada

Para obter detalhes sobre os resultados de análise discriminante:

2. Clique na guia Avançado.
3. Clique no botão "Ativar no navegador externo" (logo abaixo da guia Modelo) para visualizar os resultados em seu navegador da Web.

Analisando a saída do uso da análise de Discriminantes para classificar os clientes de telecomunicações

Análise Discriminante Stepwise

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	Age in years	1.000	1.000	7.521	.978
	Marital status	1.000	1.000	3.500	.990
	Years at current address	1.000	1.000	8.433	.975
	Household income in thousands	1.000	1.000	6.689	.980
	Level of education	1.000	1.000	61.454	.844
	Years with current employer	1.000	1.000	16.976	.951
	Retired	1.000	1.000	3.005	.991
	Gender	1.000	1.000	.373	.999
	Number of people in household	1.000	1.000	3.976	.988
1	Age in years	.980	.980	6.125	.829
	Marital status	.999	.999	3.803	.834
	Years at current address	.983	.983	8.487	.823
	Household income in thousands	.989	.989	6.022	.829
	Years with current employer	.953	.953	14.933	.807
	Retired	.992	.992	1.432	.840
	Gender	1.000	1.000	.358	.843
	Number of people in household	1.000	1.000	3.967	.834
	Age in years	.563	.548	.352	.807
	Marital status	.999	.952	3.903	.798

Figura 273. Variáveis não presentes na análise

Quando você tem um monte de preditores, o método stepwise pode ser útil selecionando automaticamente as variáveis "melhores" para usar no modelo. O método stepwise começa com um

modelo que não inclui nenhum dos preditores. Em cada etapa, o preditor com o maior valor *F to Enter* que excede os critérios de entrada (por padrão, 3.84) é incluído no modelo.

Todas as variáveis deixadas de fora da análise na última etapa têm valores *F para Inserir* menores que 3.84, portanto, não mais serão incluídas

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	Level of education	1.000	61.454	
2	Level of education	.953	59.108	.951
	Years with current employer	.953	14.933	.844
3	Level of education	.951	60.046	.940
	Years with current employer	.934	15.824	.834
	Number of people in household	.979	4.841	.807

Figura 274. Variáveis na análise

Esta tabela exibe estatísticas para as variáveis que estão na análise em cada etapa. *Tolerância* é a proporção de uma variância de uma variável não contabilizadas por outras variáveis independentes na equação. Uma variável com tolerância muito baixa contribui com pouca informação para um modelo e pode causar problemas computacionais.

Valores de *F para Remover* são úteis para descrever o que acontece se uma variável for removida do modelo atual (dado que as outras variáveis permanecem). *F to Remover* para a variável digitar é o mesmo que *F to Enter* na etapa anterior (mostrada na tabela *Variáveis Not in the Analysis*).

Uma nota de cautela relativa aos métodos sábios

Os métodos sábios são convenientes, mas têm suas limitações. Esteja ciente de que, porque os métodos stepwise selecionam modelos baseados unicamente no mérito estatístico, ele pode escolher preditores que não tenham *significância prática*. Se você tem alguma experiência com os dados e tem expectativas sobre quais os preditores são importantes, você deve usar esse conhecimento e eschiclete métodos estenteados. Se, no entanto, você tiver muitos preditores e nenhuma ideia por onde começar, executar uma análise estandarada e ajustar o modelo selecionado é melhor do que nenhum modelo em nada.

Modelo de verificação fit

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.198[a]	80.2	80.2	.407
2	.048[a]	19.4	99.6	.214
3	.001[a]	.4	100.0	.031

a.First 3 canonical discriminant functions were used in the analysis.

Figura 275. Valores próprios

Quase toda a variância explicada pelo modelo deve-se às duas primeiras funções discriminatórias. Três funções são próprias automaticamente, mas devido ao seu minuscule eigenvalue, você pode razoavelmente ignorar o terceiro.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 3	.796	227.345	9	<.001
2 through 3	.953	47.486	4	<.001
3	.999	.929	1	.335

Figura 276. Lambda de Wilks

Wilks ' lambda concorda que apenas as duas primeiras funções são úteis. Para cada conjunto de funções, este testa a hipótese de que os meios das funções listadas são iguais entre os grupos. O teste da função 3 tem um valor de significado maior que 0.10, portanto, essa função contribui pouco para o modelo.

Matriz de estrutura

Structure Matrix

	Function		
	1	2	3
Level of education	.966*	-.090	-.244
Years with current employer	-.182	.964*	-.193
Age in years[b]	-.162	.598*	-.285
Household income in thousands[b]	.109	.514*	-.190
Years at current address[b]	-.151	.394*	-.214
Retired[b]	-.108	.230*	-.137
Gender[b]	.008	.054*	.009
Number of people in household	.232	.097	.968*
Marital status[b]	.132	.134	.600*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*.Largest absolute correlation between each variable and any discriminant function

b.This variable not used in the analysis.

Figura 277. Matriz de estrutura

Quando há mais de uma função discriminante, um asterisco (*) marca a maior correlação absoluta de cada variável com uma das funções canônicas. Dentro de cada função, essas variáveis marcadas são então ordenadas pelo tamanho da correlação.

- *O nível de educação* é mais fortemente correlacionado com a primeira função, e é a única variável mais fortemente correlacionada com esta função.
- *Anos com o empregador atual*, *Idade em anos*, *Renda familiar em milhares*, *Anos em endereço atual*, *Recansado*, e *Gender* estão mais fortemente correlacionados com a segunda função, embora *Gender* e *Recansados* sejam mais fracamente correlacionados do que os outros. As outras variáveis marcam esta função como uma função de "estabilidade".
- *Número de pessoas em família* e *estado Civil* estão mais fortemente correlacionadas com a terceira função discriminante, mas esta é uma função inútil, portanto, estes são quase inúteis preditores.

Mapa territorial

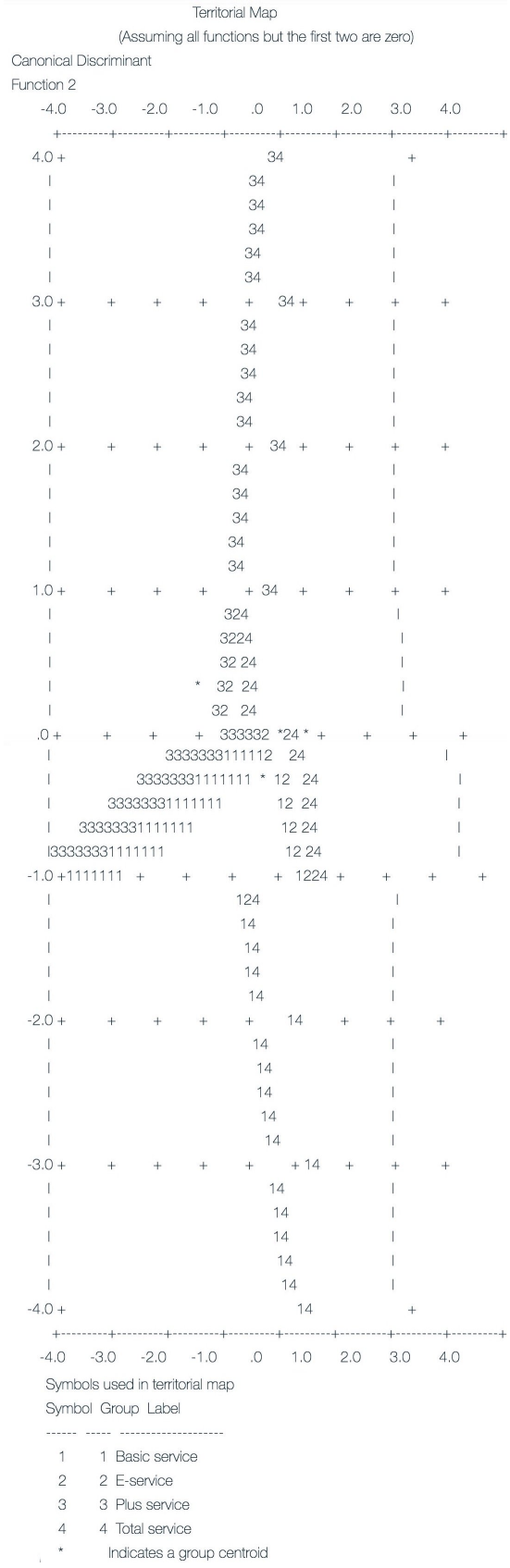


Figura 278. Mapa territorial

O mapa territorial ajuda você a estudar as relações entre os grupos e as funções discriminatórias. Combinado com os resultados matriciais da estrutura, ele dá uma interpretação gráfica da relação entre preditores e grupos. A primeira função, mostrada no eixo horizontal, separa o grupo 4 (*Total service* clientes) dos demais. Uma vez que o *Level of education* é fortemente correlacionado com a primeira função, isto sugere que seus clientes *Total service* são, em geral, os mais altamente educados. A segunda função separa os grupos 1 e 3 (*Serviço básico* e os clientes do *Plus service*). Os clientes do *Plus service* tendem a ter trabalhado por mais tempo e são mais antigos do que os clientes do *Serviço básico* . Os clientes *E-service* não são separados bem dos outros, embora o mapa sugira que eles tendem a ser bem educados com uma quantidade moderada de experiência de trabalho.

Em geral, a proximidade entre os centroides do grupo, marcados com asteriscos (*), para as linhas territoriais sugere que a separação entre todos os grupos não é muito forte.

Apenas as duas primeiras funções discriminatórias são plotadas, mas uma vez que a terceira função foi encontrada bastante insignificante, o mapa territorial oferece uma visão abrangente do modelo discriminante.

Resultados da classificação

Classification Results[a]

		Customer category	Predicted Group Membership				Total
			Basic service	E-service	Plus service	Total service	
Original	Count	Basic service	125	11	61	69	266
		E-service	49	15	58	95	217
		Plus service	102	14	112	53	281
		Total service	40	16	37	143	236
	%	Basic service	47.0	4.1	22.9	25.9	100.0
		E-service	22.6	6.9	26.7	43.8	100.0
		Plus service	36.3	5.0	39.9	18.9	100.0
		Total service	16.9	6.8	15.7	60.6	100.0

a. 39.5% of original grouped cases correctly classified.

Figura 279. Resultados da classificação

Da lambda de Wilks, você sabe que seu modelo está se saindo melhor do que adivinhação, mas você precisa recorrer aos resultados da classificação para determinar o quanto melhor. Considerando os dados observados, o modelo "null" (ou seja, um sem preditores) classificaria todos os clientes para o grupo modal, *Plus service*. Assim, o modelo nulo estaria correto 281/1000 = 28.1% do tempo. Seu modelo obtém 11.4% mais ou 39.5% dos clientes. Em particular, seu modelo se sobressai na identificação de clientes *Total service* . No entanto, faz um trabalho excepcionalmente pobre de classificação de clientes *E-service* . Você pode precisar encontrar outro preditor a fim de separar esses clientes.

Resumo

Você criou um modelo discriminante que classifica os clientes em um dos quatro grupos predefinidos de "uso de serviço", com base em informações demográficas de cada cliente. Utilizando a matriz de estrutura e o mapa territorial, você identificou quais variáveis são mais úteis para segmentar sua base de clientes. Por último, os resultados de classificação mostram que o modelo faz mal em classificar os clientes *E-service* . Mais pesquisas são necessárias para determinar outra variável de preditor que melhor classifique esses clientes, mas dependendo do que você está procurando prever, o modelo pode estar perfeitamente adequado para suas necessidades. Por exemplo, se você não está preocupado em identificar os clientes *E-service* o modelo pode ser preciso o suficiente para você. Este pode ser o caso em que o E-service é um líder deficitária que traz pouco lucro. Se, por exemplo, o seu maior retorno sobre investimento vier de clientes *Plus service* ou *Total service* , o modelo pode te dar as informações que você precisa.

Observe também que esses resultados são baseados apenas nos dados de treinamento. Para avaliar o quão bem o modelo se generaliza para outros dados, você pode usar um nó Partition para realizar um subconjunto de registros para fins de teste e validação.

As explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no IBM SPSS Modelador Guia de Algoritmos. Isso está disponível a partir do diretório *Documentação* do disco de instalação.

Capítulo 22. Analisando dados de sobrevivência censurados (Generalized Linear Models)

Ao analisar dados de sobrevivência com censura de intervalo, ou seja, quando o tempo exato do evento de interesse não é conhecido, mas sabe-se apenas que ocorreu em um determinado intervalo, a aplicação do modelo Cox aos riscos de eventos em intervalos resulta em um modelo de regressão logarítmica complementar.

As informações parciais de um estudo projetado para comparar a eficácia de duas terapias para evitar a recorrência de úlceras são coletadas em *ulcer_recurrence.sav* Este dataset foi apresentado e analisado em outro lugar[1]. Utilizando modelos lineares generalizados, é possível replicar os resultados para os modelos de regressão log-log complementares.

Este exemplo usa o fluxo denominado *ulcer_genlin.str*, que faz referência ao arquivo de dados *ulcer_recurrence.sav* O arquivo de dados está na pasta *Demos* e o arquivo stream está na subpasta *streams* .

Criando o Stream

1. Inclua um nó de origem do Arquivo de Estatísticas apontando para *ulcer_recurrence.sav* na pasta *Demos* .

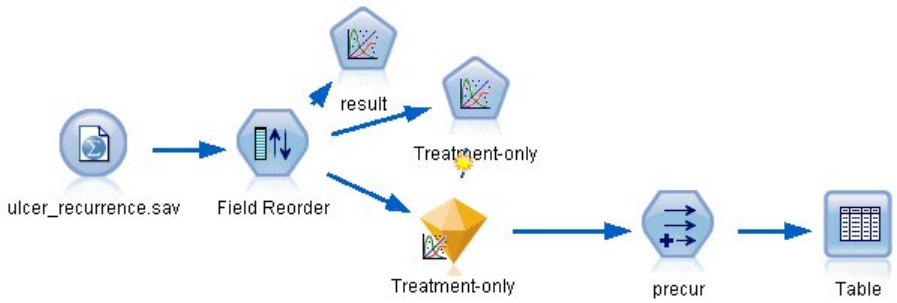


Figura 280. Fluxo de amostra para prever recorrência de úlcera

2. Na guia Filtro do nó de origem, filtre-se *id* e *time*.

[1] Collett, D. 2003. *Modelling survival data in medical research*, 2ª ed. Boca Raton: Chapman & Hall/CRC.

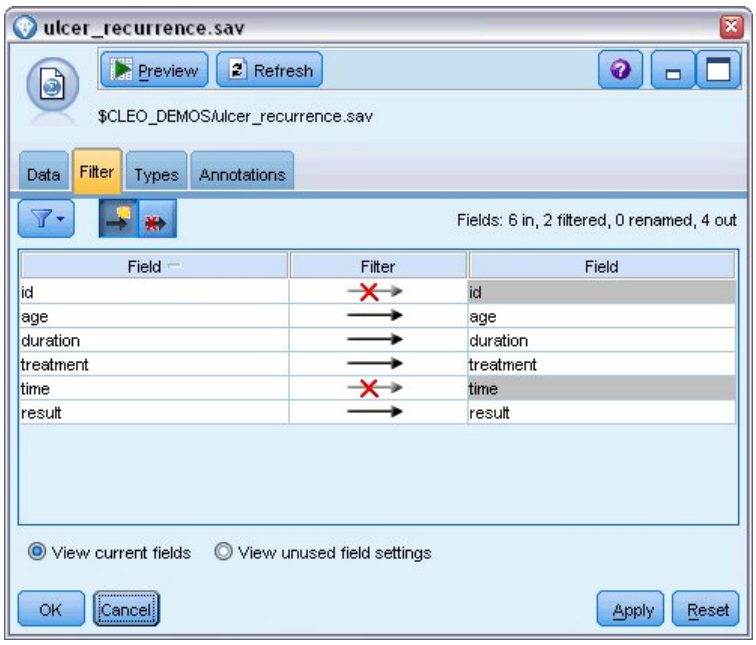


Figura 281. Filtrar campos indesejados

3. Na guia Tipos do nó de origem, configure a função para o campo *result* para **Target** e configure seu nível de medição para **Flag**. Um resultado de 1 indica que a úlcera se repetiu. Todos os outros campos devem ter seu papel configurado como **Entrada**.
4. Clique em **Valores de leitura** para instanciar os dados.

Figura 282. Configurando função de campo

5. Adiciam um nó de Reordem de Campo e especificam *duração*, *tratamento*e *idade* como a ordem de entradas. Isso determina a ordem em que os campos estão inseridos no modelo e irá ajudá-lo a tentar replicar os resultados de Collett.

Figura 283. Reordenando campos para que eles sejam inseridos no modelo conforme desejado

6. Conecte um nó GenLin ao nó de origem; no nó GenLin , clique na guia **Modelo** .
7. Selecione **Primeiro (Lowest)** como a categoria de referência para o destino. Isso indica que a segunda categoria é o evento de interesse, e seu efeito sobre o modelo está na interpretação de estimativas de parâmetros. Um preditor contínuo com coeficiente positivo indica aumento da probabilidade de recorrência com valores crescentes do preditor; categorias de um preditor nominal com coeficientes maiores indicam aumento de probabilidade de recorrência com relação a outras categorias do conjunto.

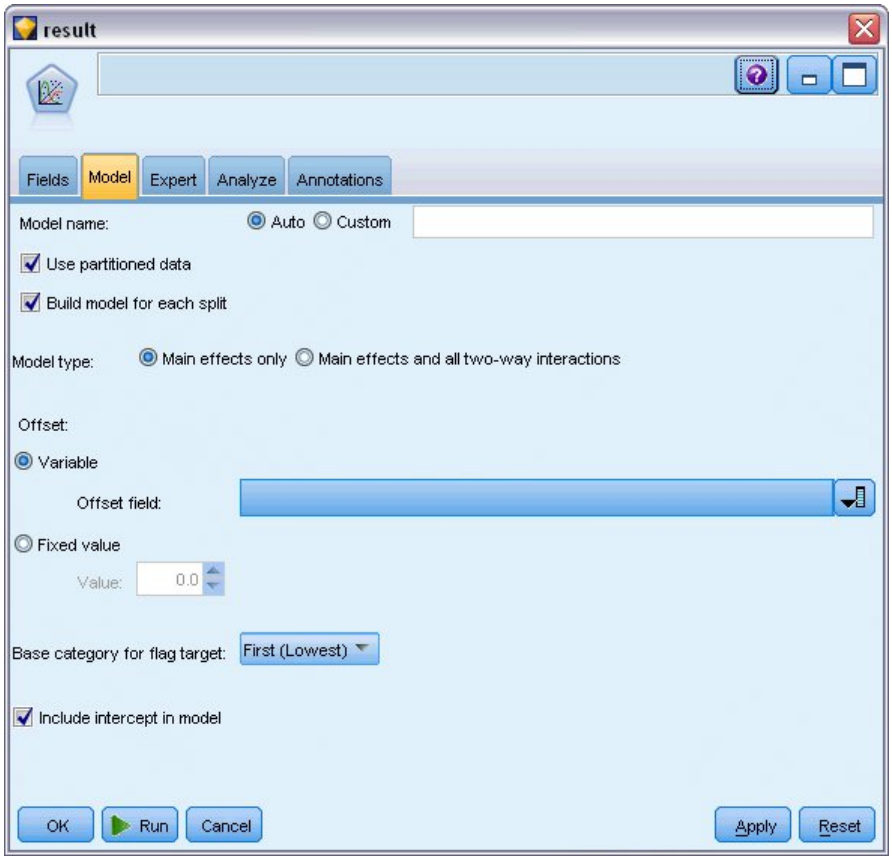


Figura 284. Escolhendo opções de modelo

8. Clique na guia **Expert** e selecione **Expert** para ativar as opções de modelagem de especialistas.
9. Selecione **Binomial** como a distribuição e **log complementar-log** como a função link.
10. Selecione **Valor fixo** como o método para estimar o parâmetro de escala e deixe o valor padrão de 1.0.
11. Selecione **Descender** como a ordem de categoria para fatores. Isso indica que a primeira categoria de cada fator será a sua categoria de referência; o efeito dessa seleção sobre o modelo está na interpretação de estimativas de parâmetros.

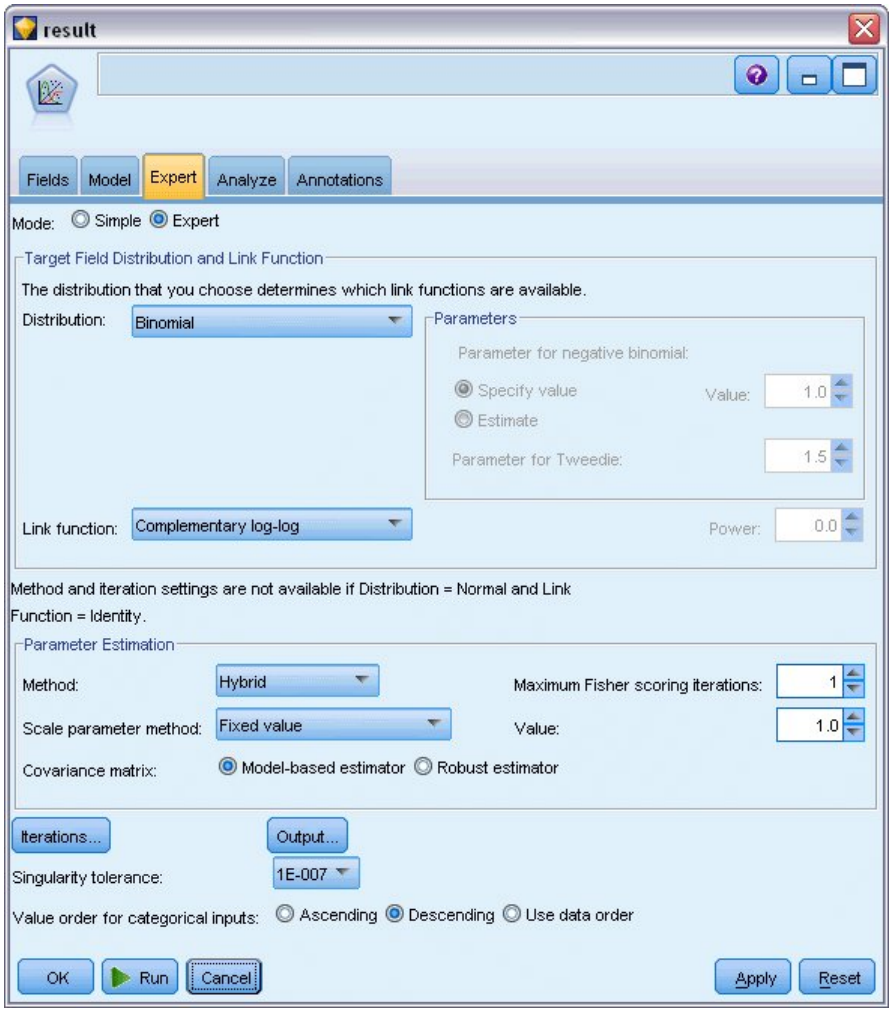


Figura 285. Escolhendo as opções de especialistas

12. Execute o fluxo para criar o nugget do modelo, que é adicionado à tela do fluxo, e também à paleta de Modelos no canto superior direito. Para visualizar os detalhes do modelo, clique com o botão direito do mouse sobre o nugget e escolha **Editar** ou **Procurar**.

Teste de efeitos do modelo

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	.536	1	.464
Age in years	.358	1	.550
Duration of disease	.003	1	.958
Treatment group	.382	1	.537

Dependent Variable: Result

Model: (Intercept), Age in years, Duration of disease, Treatment group

Figura 286. Ensaios de efeitos de modelo para modelo de efeitos principais

Nenhum dos efeitos de modelo é estatisticamente significativo; no entanto, quaisquer diferenças observáveis nos efeitos de tratamento são de interesse clínico, por isso vamos encaixar um modelo reduzido com apenas o tratamento como um termo de modelo.

Encaixando o modelo de tratamento apenas

1. Na guia Campos do nó GenLin , clique em **Usar configurações customizadas**.
2. Selecione *resultado* como o destino.
3. Selecione *tratamento* como a entrada exclusiva.

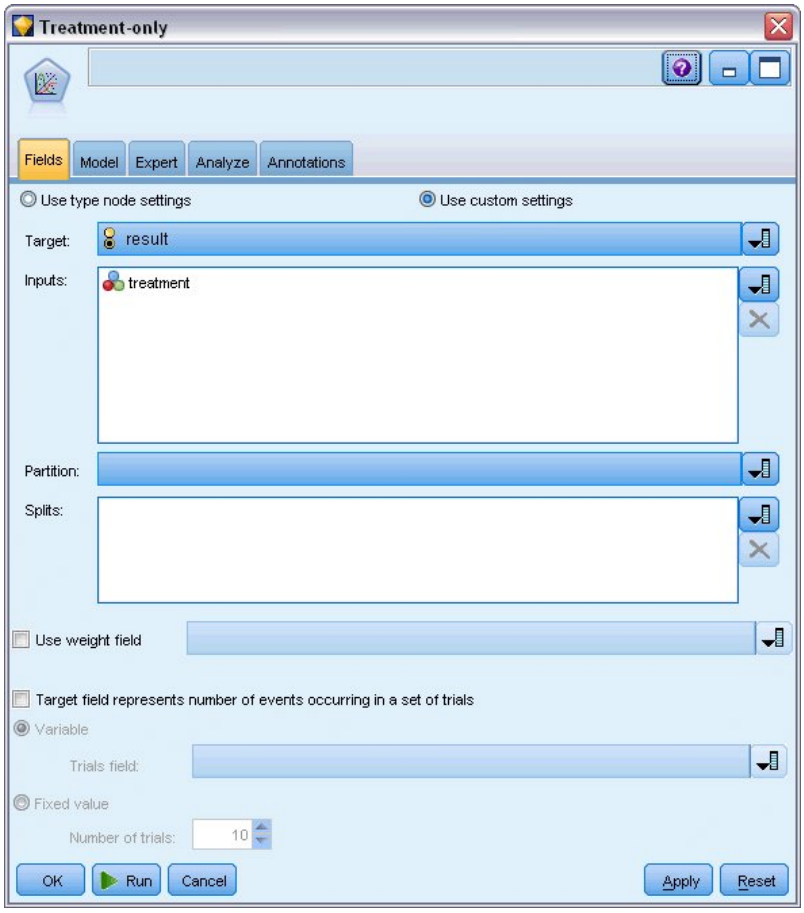


Figura 287. Escolhendo opções de campo

4. Execute o fluxo e abra o nugget de modelo resultante.

No modelo nugget, selecione a guia **Avançado** e role para a parte inferior.

Estimativas de parâmetro

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-1.442	.5012	-2.425	-.460	8.282	1	.004
[Treatment group=1]	.378	.6288	-.855	1.610	.361	1	.548
[Treatment group=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result

Model: (Intercept), Treatment group

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Figura 288. Estimativas de parâmetros para o modelo de tratamento

O efeito de tratamento (a diferença do predictor linear entre os dois níveis de tratamento; ou seja, o coeficiente para *[treatment=1]*) ainda não é estatisticamente significativo, mas apenas sugestivo que o tratamento *A [treatment=0]* pode ser melhor do que *B [treatment=1]* porque a estimativa de parâmetro para o tratamento *B* é maior do que a para *A*, estando assim associada a uma probabilidade aumentada

de recorrência nos primeiros 12 meses. O preditor linear, (intercepto + efeito de tratamento) é uma estimativa de log (–log (1–P (repetição$_{12,t}$)), em que P (repetição$_{12,t}$) é a probabilidade de recorrência em 12 meses para o tratamento t (=*A* ou *B*). Essas probabilidades previstas são geradas para cada observação no dataset.

Probabilidades de recorrência e sobrevivência previstas

Figura 289. Opções de configurações do nó deriv

1. Para cada paciente, o modelo marca o resultado previsto e a probabilidade de esse resultado previsto. A fim de ver as probabilidades de recorrência previstas, copie o modelo gerado para a paleta e anexe um nó de Derivação.
2. Na guia Configurações, digite `precur` como o campo derivado.
3. Opte por derivá-lo como **Condicionado**.
4. Clique no botão calculadora para abrir o Expression Builder para a condição **If** .

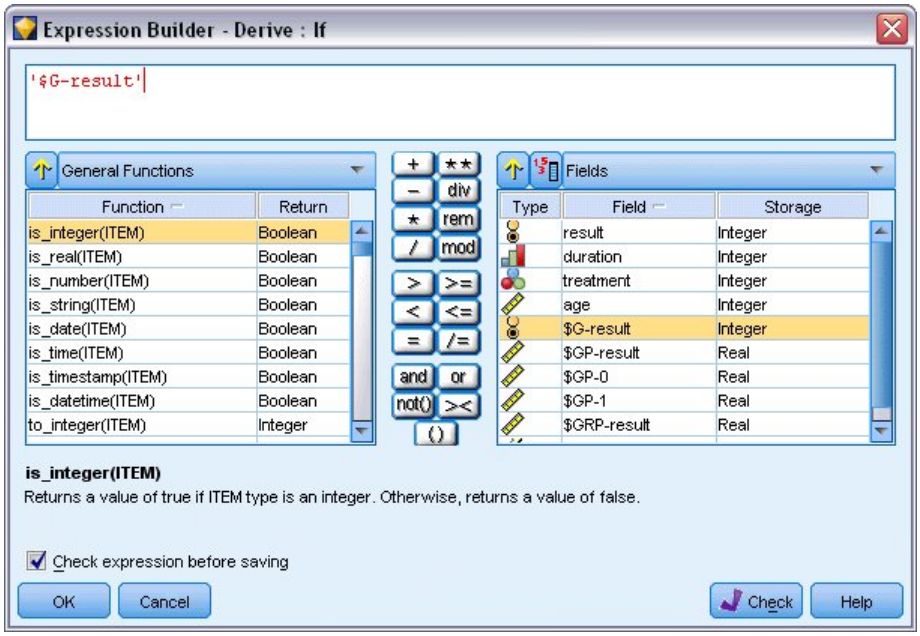


Figura 290. Derivar nó: Expression Builder para Se condição

5. Insira o campo *$G-result* na expressão.
6. Clique em **OK**.

O campo derivar *precur* irá tirar o valor da expressão **Then** quando *$G-result* é igual a 1 e o valor da expressão **Else** quando for 0.

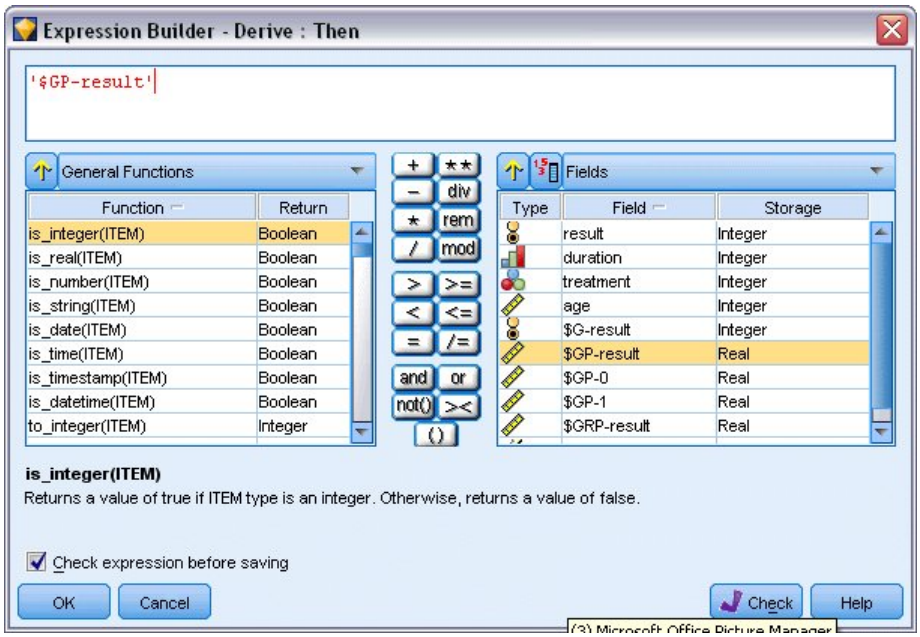


Figura 291. Derivar nó: Expression Builder para Então expressão

7. Clique no botão calculadora para abrir o Expression Builder para a expressão **Then** .
8. Insira o campo *$GP-result* na expressão.
9. Clique em **OK**.

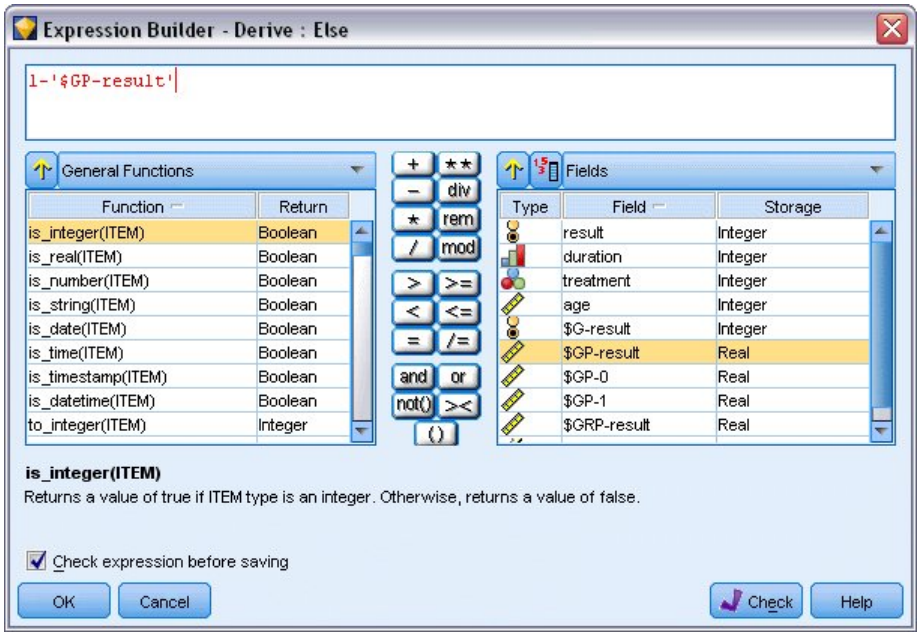
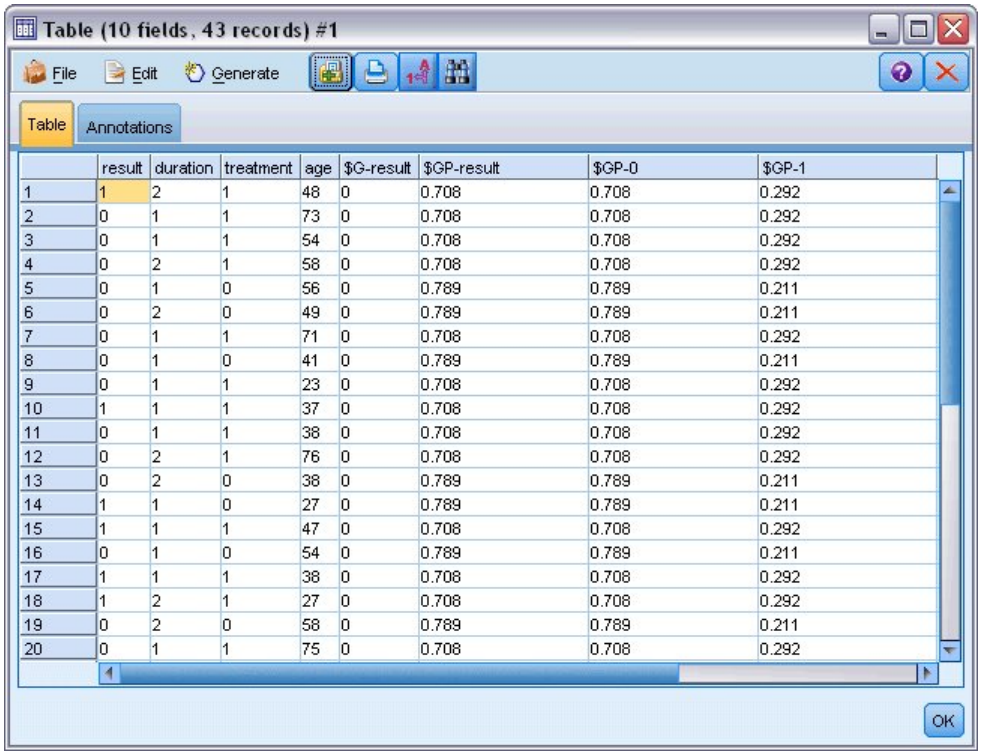


Figura 292. Deriva do nó: Expression Builder para expressão Else

10. Clique no botão calculadora para abrir o Expression Builder para a expressão **Else** .
11. Digite 1 - na expressão e insira o campo *$GP-result* na expressão.
12. Clique em **OK**.

Figura 293. Opções de configurações do nó deriv

13. Conecte um nó de tabela ao nó do Derivar e execute-o.



	result	duration	treatment	age	$G-result	$GP-result	$GP-0	$GP-1
1	1	2	1	48	0	0.708	0.708	0.292
2	0	1	1	73	0	0.708	0.708	0.292
3	0	1	1	54	0	0.708	0.708	0.292
4	0	2	1	58	0	0.708	0.708	0.292
5	0	1	0	56	0	0.789	0.789	0.211
6	0	2	0	49	0	0.789	0.789	0.211
7	0	1	1	71	0	0.708	0.708	0.292
8	0	1	0	41	0	0.789	0.789	0.211
9	0	1	1	23	0	0.708	0.708	0.292
10	1	1	1	37	0	0.708	0.708	0.292
11	0	1	1	38	0	0.708	0.708	0.292
12	0	2	1	76	0	0.708	0.708	0.292
13	0	2	0	38	0	0.789	0.789	0.211
14	1	1	0	27	0	0.789	0.789	0.211
15	1	1	1	47	0	0.708	0.708	0.292
16	0	1	0	54	0	0.789	0.789	0.211
17	1	1	1	38	0	0.708	0.708	0.292
18	1	2	1	27	0	0.708	0.708	0.292
19	0	2	0	58	0	0.789	0.789	0.211
20	0	1	1	75	0	0.708	0.708	0.292

Figura 294. Probabilidades preditas

Há uma probabilidade estimada de 0.211 de que os pacientes designados ao tratamento *A* terão uma recorrência nos primeiros 12 meses; 0.292 para o tratamento *B*. Observe que 1 −P (repita$_{12,\,t}$) é a probabilidade de sobrevivência em 12 meses, o que pode ser de mais interesse para analistas de sobrevivência.

Modelagem da probabilidade de recorrência por período

Um problema com o modelo tal como está é que ele ignora as informações recolhidas no primeiro exame; ou seja, que muitos pacientes não experimentaram uma recorrência nos primeiros seis meses. Um modelo "melhor" modelaria uma resposta binária que registra se o evento ocorreu ou não durante cada intervalo. O ajuste desse modelo requer uma reconstrução do conjunto de dados original, que pode ser localizado em *ulcer_recurrence_recoded.sav* Este arquivo contém duas variáveis adicionais:

- *Período*, que registra se o caso corresponde ao primeiro período de exame ou ao segundo.
- *Resultado por período*, que registra se houve recorrência para o paciente determinado durante o período determinado.

Cada caso original (paciente) contribui com um caso por intervalo em que permanece no conjunto de riscos. Assim, por exemplo, o paciente 1 contribui com dois casos; um para o primeiro período de exame em que nenhuma recorrência ocorreu, e um para o segundo período de exame, no qual foi registrada uma recorrência. O paciente 10, por outro lado, contribui com um único caso porque uma recorrência foi registrada no primeiro período. Os pacientes 16, 28 e 34 desistam do estudo após seis meses e, assim, contribuem apenas com um único caso para o novo dataset.

1. Inclua um nó de origem do Arquivo de Estatísticas apontando para *ulcer_recurrence_recoded.sav* na pasta *Demos* .

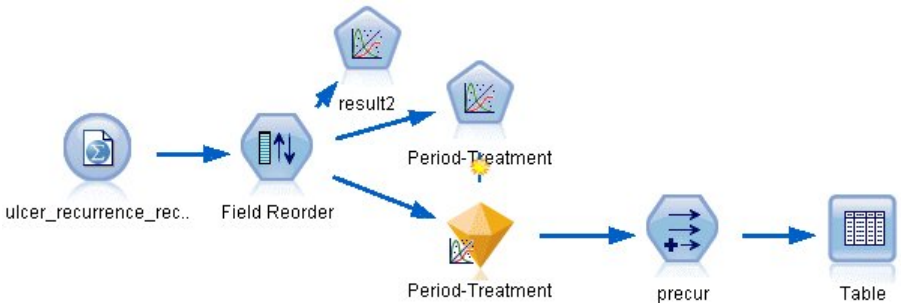


Figura 295. Fluxo de amostra para prever recorrência de úlcera

2. Na guia Filtro do nó de origem, filtre-se *id*, *timee result*.

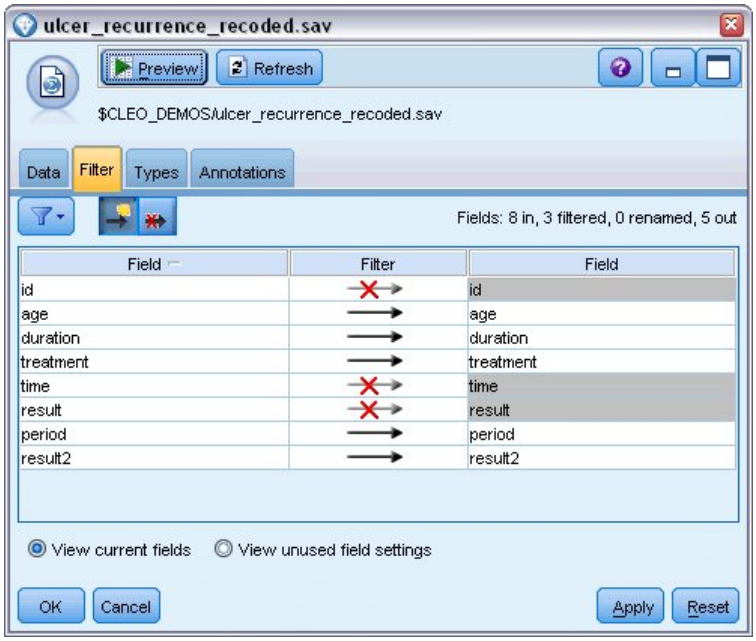


Figura 296. Filtrar campos indesejados

3. Na guia Tipos do nó de origem, configure a função para o campo *result2* para **Destino** e configure seu nível de medição para **Bandeira**. Todos os outros campos devem ter seu papel configurado como **Entrada**.

Figura 297. Configurando função de campo

4. Adiciam um nó de Reordem de Campo e especificam *período*, *duração*, *tratamento*e *idade* como a ordem de entradas. Fazer *período* a primeira entrada (e não incluir o termo de interceptação no modelo) permitirá que você se encaixe em um conjunto completo de variáveis dummy para capturar os efeitos do período.

Figura 298. Reordenando campos para que eles sejam inseridos no modelo conforme desejado

5. No nó GenLin , clique na guia **Modelo** .

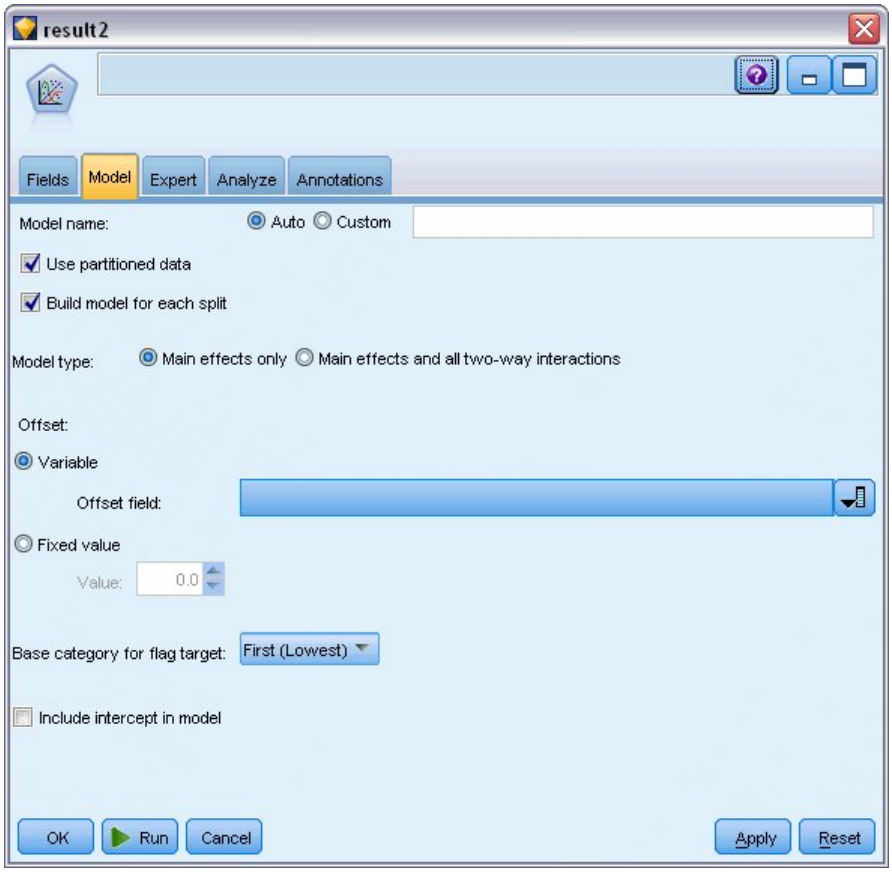


Figura 299. Escolhendo opções de modelo

6. Selecione **Primeiro (Lowest)** como a categoria de referência para o destino. Isso indica que a segunda categoria é o evento de interesse, e seu efeito sobre o modelo está na interpretação de estimativas de parâmetros.
7. Desmarque **Include intercepto no modelo**.
8. Clique na guia **Expert** e selecione **Expert** para ativar as opções de modelagem de especialistas.

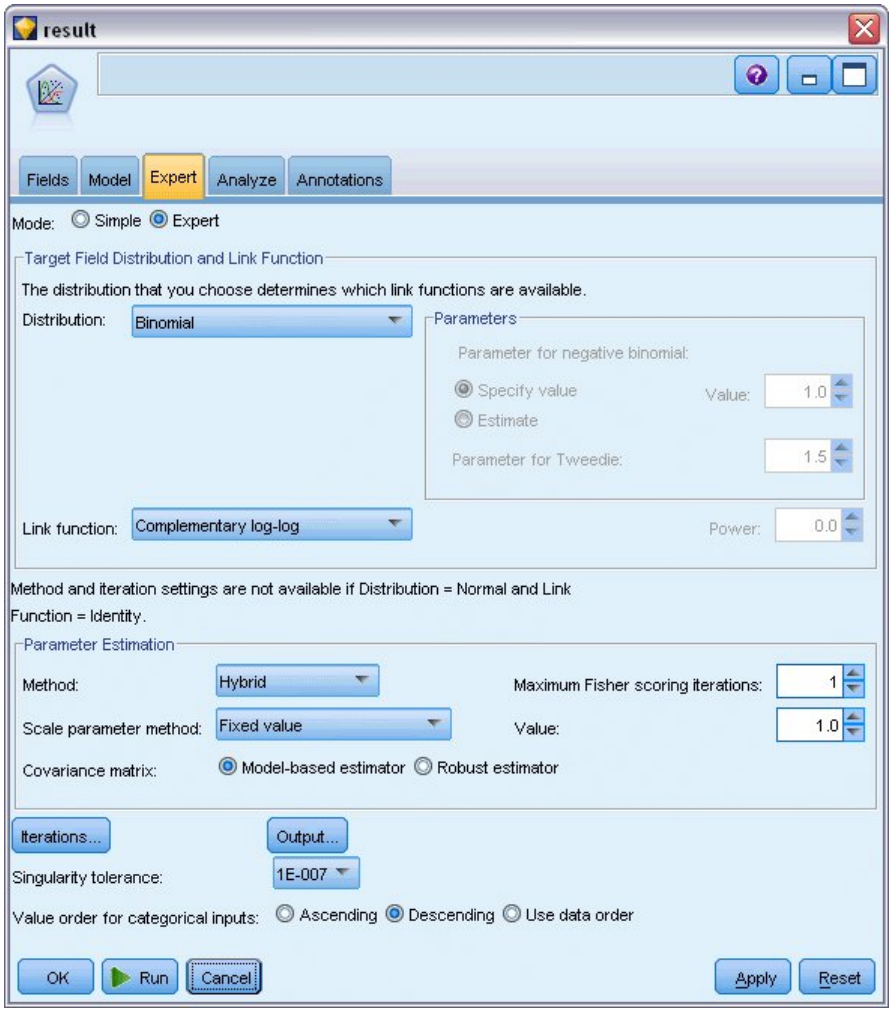


Figura 300. Escolhendo as opções de especialistas

9. Selecione **Binomial** como a distribuição e **log complementar-log** como a função link.
10. Selecione **Valor fixo** como o método para estimar o parâmetro de escala e deixe o valor padrão de 1.0.
11. Selecione **Descender** como a ordem de categoria para fatores. Isso indica que a primeira categoria de cada fator será a sua categoria de referência; o efeito dessa seleção sobre o modelo está na interpretação de estimativas de parâmetros.
12. Execute o fluxo para criar o nugget do modelo, que é adicionado à tela do fluxo, e também à paleta de Modelos no canto superior direito. Para visualizar os detalhes do modelo, clique com o botão direito do mouse sobre o nugget e escolha **Editar** ou **Procurar**.

Teste de efeitos do modelo

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
Period	.464	1	.496
Age in years	.314	1	.575
Duration of disease	.000	1	.988
Treatment group	.117	1	.732

Dependent Variable: Result by period

Model: Period, Age in years, Duration of disease, Treatment group

Figura 301. Ensaios de efeitos de modelo para modelo de efeitos principais

Nenhum dos efeitos de modelo é estatisticamente significativo; no entanto, quaisquer diferenças observáveis no período e efeitos de tratamento são de interesse clínico, por isso vamos encaixar um modelo reduzido com apenas esses termos de modelo.

Encaixe o modelo reduzido

1. Na guia Campos do nó GenLin , clique em **Usar configurações customizadas**.
2. Selecione *result2* como o alvo.
3. Selecione *período* e *tratamento* como as entradas.

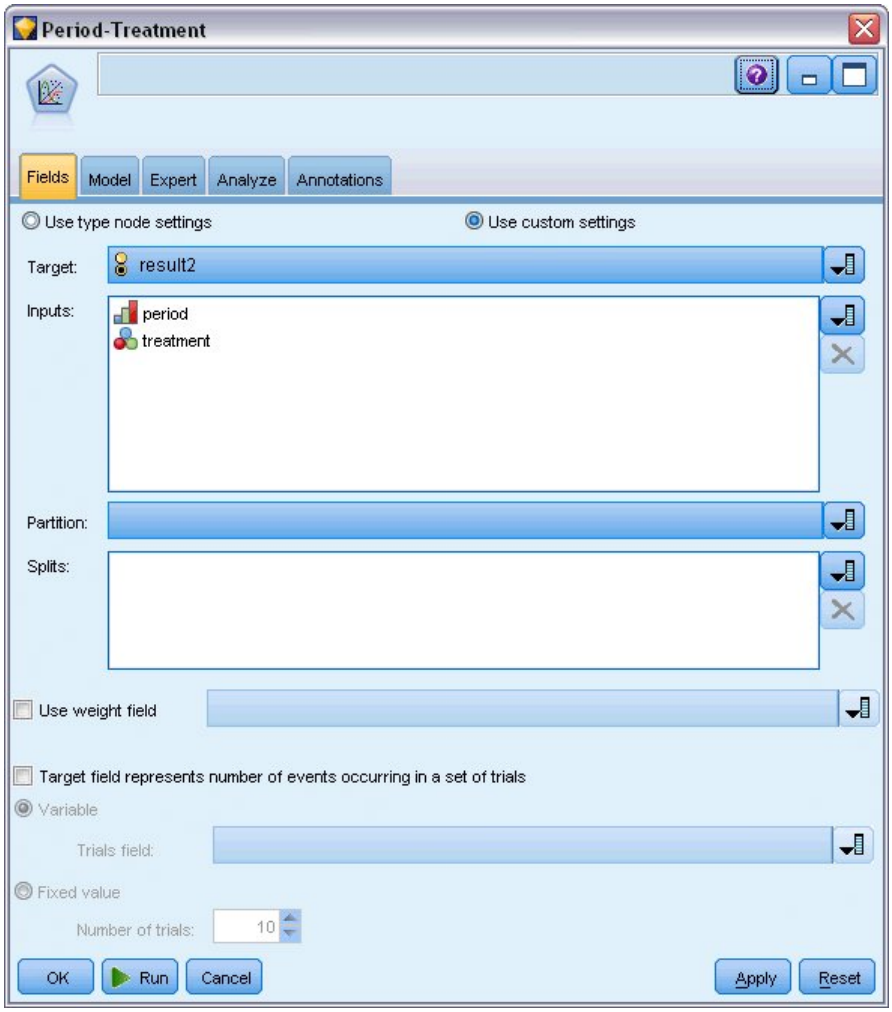


Figura 302. Escolhendo opções de campo

4. Execute o nó e navegue pelo modelo gerado e, em seguida, copie o modelo gerado para a paleta, anexe um nó de tabela e execute-o.

Estimativas de parâmetro

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
[Period=2]	-1.794	.5792	-2.929	-.659	9.597	1	.002
[Period=1]	-2.206	.5912	-3.365	-1.047	13.926	1	.000
[Treatment group=1]	.195	.6279	-1.035	1.426	.097	1	.756
[Treatment group=0]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: Result by period
Model: Period, Treatment group

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Figura 303. Estimativas de parâmetros para o modelo de tratamento

O efeito de tratamento ainda não é estatisticamente significativo mas apenas sugestivo que o tratamento *A* pode ser melhor do que o *B* porque a estimativa de parâmetro para o tratamento *B* está associada a uma probabilidade aumentada de recorrência nos primeiros 12 meses. Os valores de período são estatisticamente significativamente diferentes a partir de 0, mas isso ocorre por causa do fato de que um termo de interceptação não está apto. O efeito de período (a diferença entre os valores do preditor linear para *[period=1]* e *[period=2]*) não é estatisticamente significante, como pode ser visto nos testes de efeitos de modelo. O predictor linear (efeito de período + efeito de tratamento) é uma estimativa de log (−log (1−P (repetição$_{p, t}$)), em que P (repetição$_{p, t}$) é a probabilidade de recorrência no período p (= 1 ou 2, representando seis meses ou 12 meses) dado tratamento t (=*A* ou *B*). Essas probabilidades previstas são geradas para cada observação no dataset.

Probabilidades de recorrência e sobrevivência previstas

Figura 304. Opções de configurações do nó deriv

1. Para cada paciente, o modelo marca o resultado previsto e a probabilidade de esse resultado previsto. A fim de ver as probabilidades de recorrência previstas, copie o modelo gerado para a paleta e anexe um nó de Derivação.
2. Na guia Configurações, digite `precur` como o campo derivado.
3. Opte por derivá-lo como **Condicionado**.
4. Clique no botão calculadora para abrir o Expression Builder para a condição **If** .

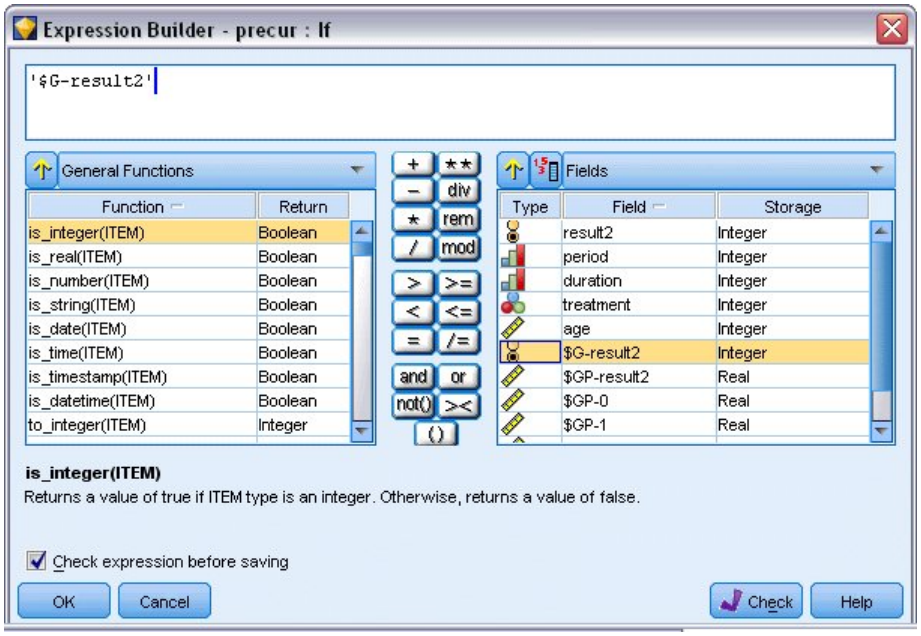


Figura 305. Derivar nó: Expression Builder para Se condição

5. Insira o campo *$G-result2* na expressão.
6. Clique em **OK**.

O campo derivar *precur* levará o valor da expressão **Em seguida** quando *$G-result2* é igual a 1 e o valor da expressão **Else** quando ele for 0.

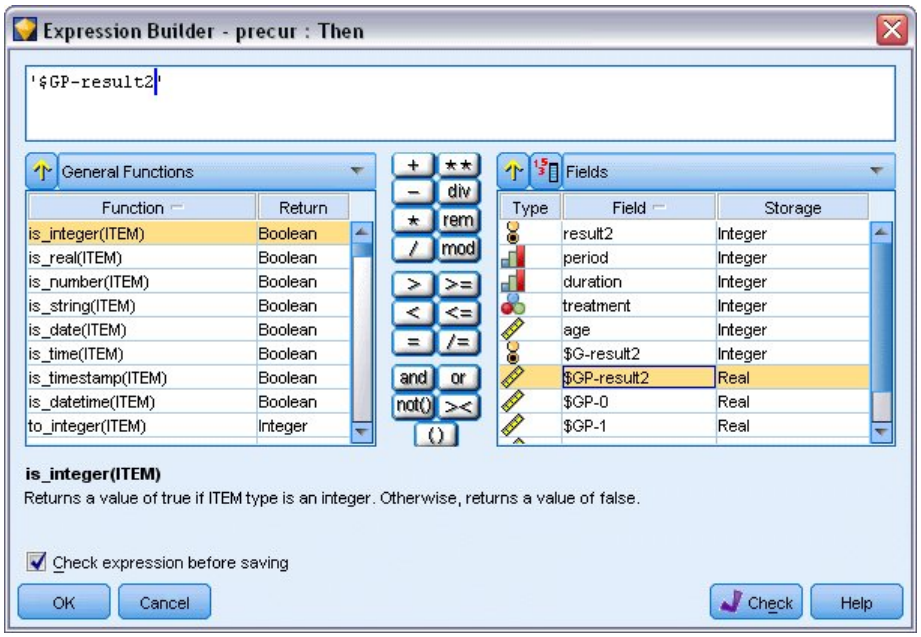


Figura 306. Derivar nó: Expression Builder para Então expressão

7. Clique no botão calculadora para abrir o Expression Builder para a expressão **Then** .
8. Insira o campo *$GP-result2* na expressão.
9. Clique em **OK**.

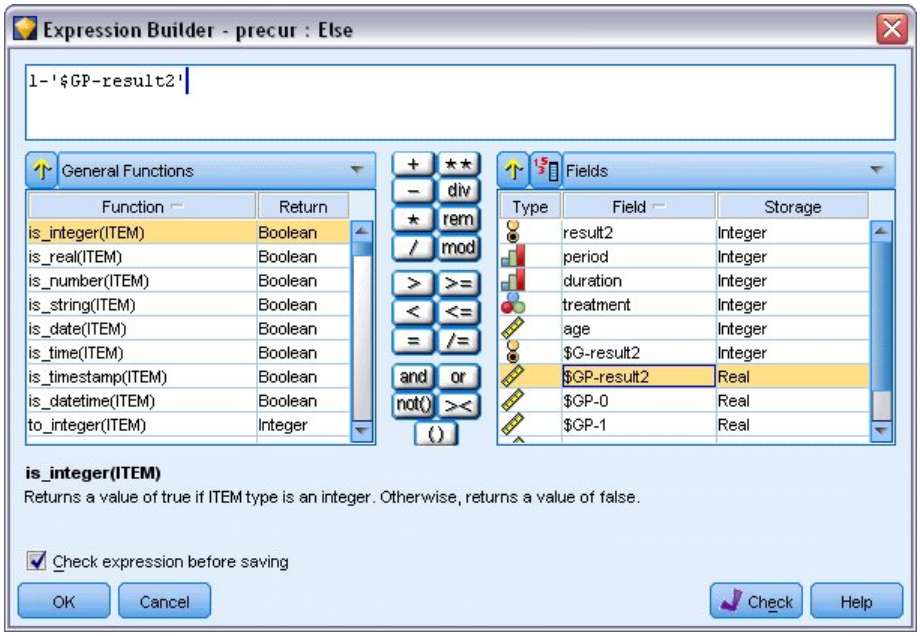


Figura 307. Deriva do nó: Expression Builder para expressão Else

10. Clique no botão calculadora para abrir o Expression Builder para a expressão **Else** .
11. Digite 1- na expressão e, em seguida, insira o campo *$GP-result2* na expressão.
12. Clique em **OK**.

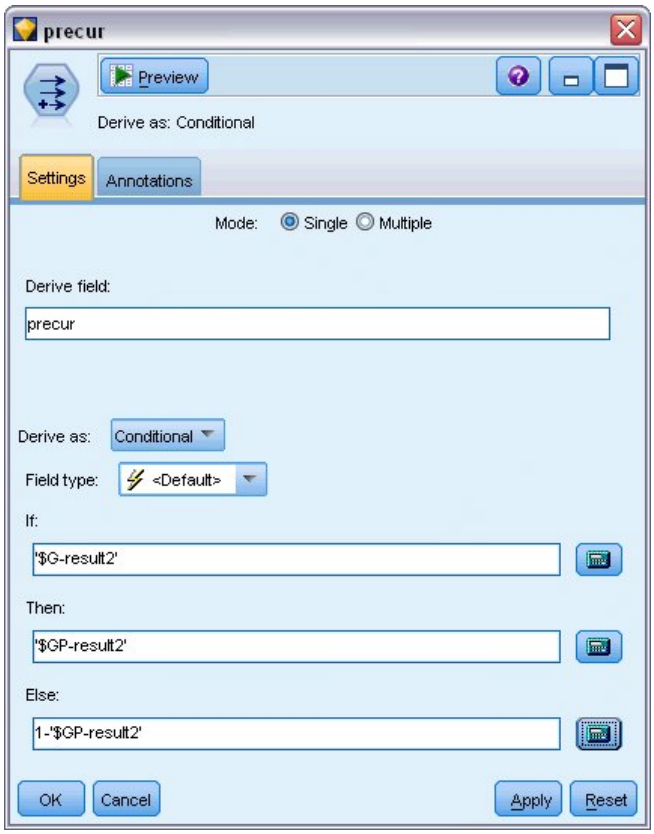
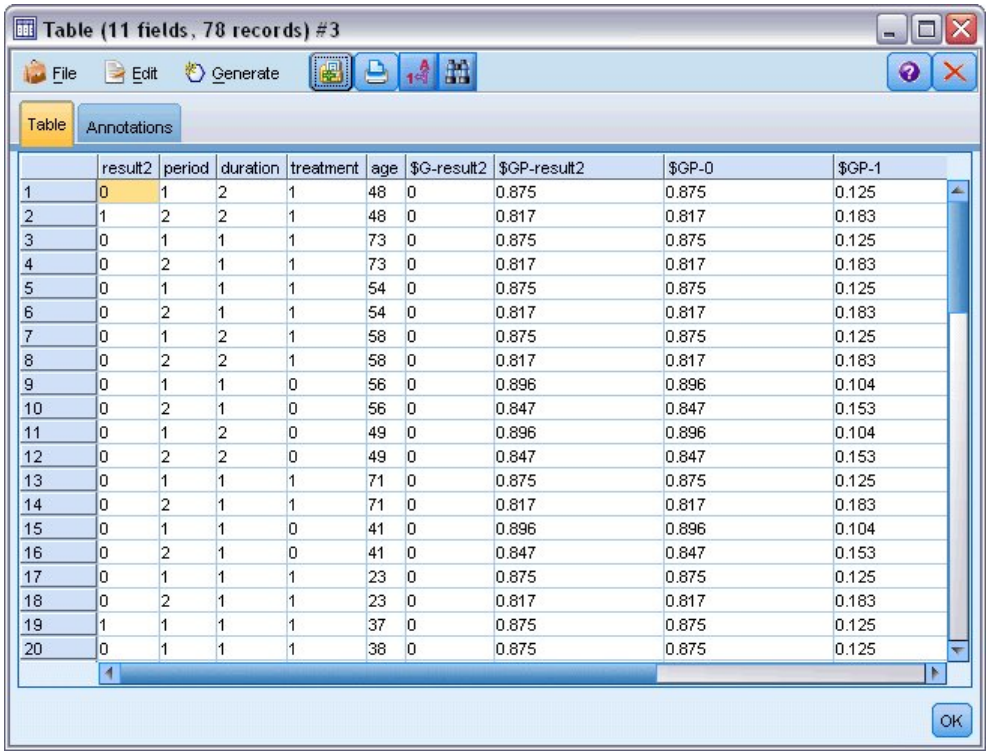


Figura 308. Opções de configurações do nó deriv

13. Conecte um nó de tabela ao nó do Derivar e execute-o.



	result2	period	duration	treatment	age	$G-result2	$GP-result2	$GP-0	$GP-1
1	0	1	2	1	48	0	0.875	0.875	0.125
2	1	2	2	1	48	0	0.817	0.817	0.183
3	0	1	1	1	73	0	0.875	0.875	0.125
4	0	2	1	1	73	0	0.817	0.817	0.183
5	0	1	1	1	54	0	0.875	0.875	0.125
6	0	2	1	1	54	0	0.817	0.817	0.183
7	0	1	2	1	58	0	0.875	0.875	0.125
8	0	2	2	1	58	0	0.817	0.817	0.183
9	0	1	1	0	56	0	0.896	0.896	0.104
10	0	2	1	0	56	0	0.847	0.847	0.153
11	0	1	2	0	49	0	0.896	0.896	0.104
12	0	2	2	0	49	0	0.847	0.847	0.153
13	0	1	1	1	71	0	0.875	0.875	0.125
14	0	2	1	1	71	0	0.817	0.817	0.183
15	0	1	1	0	41	0	0.896	0.896	0.104
16	0	2	1	0	41	0	0.847	0.847	0.153
17	0	1	1	1	23	0	0.875	0.875	0.125
18	0	2	1	1	23	0	0.817	0.817	0.183
19	1	1	1	1	37	0	0.875	0.875	0.125
20	0	1	1	1	38	0	0.875	0.875	0.125

Figura 309. Probabilidades preditas

Tabela 3. Probabilidades de recorrência estimadas

Tratamento	**6 meses**	**12 meses**
A	0.104	0.153
B	0.125	0.183

A partir das probabilidades de recorrência estimadas, a probabilidade de sobrevivência através de 12 meses pode ser estimada como $1 - (P(\text{recorrentes}_{1,t}) + P(\text{recorrentes}_{2,t}) \times (1 - P(\text{recorrentes}_{1,t})))$; assim, para cada tratamento:

A : 1 - (0.104 + 0.153 * 0.896) = 0.759

B : 1 - (0.125 + 0.183 * 0.875) = 0.715

que novamente mostra um suporte não estatisticamente significativo para *A* como o melhor tratamento.

Resumo

Utilizando Modelos Lineares Generalizados, você se encaixou em uma série de modelos de regressão log-log complementares para dados de sobrevivência censurada por intervalos. Embora haja algum suporte para a escolha do tratamento *A*, alcançar um resultado estatisticamente significativo pode exigir um estudo maior. No entanto, há mais algumas avenidas para explorar com os dados existentes.

- Pode valer a pena reajustá-lo o modelo com efeitos de interação, particularmente entre *Período* e *Grupo de Tratamento*.

Explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no *IBM SPSS Modelador Algorithms Guide*.

Procedimentos Relacionados

O Procedimento Modelos Lineares Generalizados é uma ferramenta poderosa para encaixar uma variedade de modelos.

- O Procedimento de Equações de Estimativa Generalizadas amplia o modelo linear generalizado para permitir medições repetidas.
- O Procedimento de Modelos Mistas Lineares permite que você se encaixe em modelos para variáveis dependentes de escala com um componente aleatório e / ou medições repetidas.

Leituras recomendadas

Veja os textos a seguir para obter mais informações sobre modelos lineares generalizados:

Cameron, A. C., e P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2ª ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., e J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., e J. A. Nelder. 1989. *Generalized Linear Models*, 2ª ed. Londres: Chapman & Hall.

Capítulo 23. Usando regressão de poisson para analisar taxas de danos do navio (Modelos Lineares Generalizados)

Um modelo linear generalizado pode ser usado para se encaixar em uma regressão de Poisson para a análise de dados de contagem. Por exemplo, um dataset apresentado e analisado em outro lugar[2]diz respeito aos danos causados por navios de carga causados por ondas. As contagens de incidentes podem ser modeladas como ocorrendo a uma taxa de Poisson dada os valores dos preditores, e o modelo resultante pode ajudá-lo a determinar quais tipos de navios são mais propensos a danos.

Este exemplo usa o fluxo *ships_genlin.str*, que faz referência ao arquivo de dados *ships.sav* O arquivo de dados está na pasta *Demos* e o arquivo stream está na subpasta *streams* .

A modelagem das contagens de células brutas pode ser enganosa nessa situação porque o *Agregado meses de serviço* varia por tipo de navio. Variáveis como esta que medem a quantidade de "exposição" ao risco são tratadas dentro do modelo linear generalizado como variáveis de deslocamento. Além disso, uma regressão de Poisson assume que o log da variável dependente é linear nos preditores. Assim, para utilizar modelos lineares generalizados para encaixar uma regressão de Poisson às taxas de acidentes, é necessário utilizar *Logaritmo de meses agregados de serviço*.

Encaixando um regressão Poisson "superdisperso"

1. Inclua um nó de origem do Arquivo de Estatísticas apontando para *ships.sav* na pasta *Demos* .

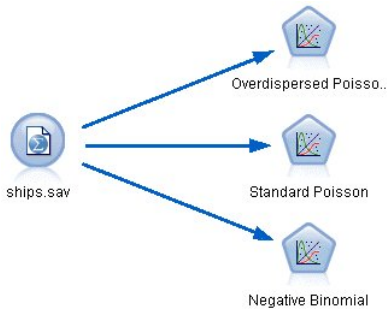


Figura 310. Fluxo de amostra para analisar taxas de danos

2. Na guia Filtro do nó de origem, exclua o campo *months_service*. Os valores transformados em log desta variável estão contidos em *log_months_service*, que serão utilizados na análise.

[2] McCullagh, P., e J. A. Nelder. 1989. *Generalized Linear Models*, 2ª ed. Londres: Chapman & Hall.

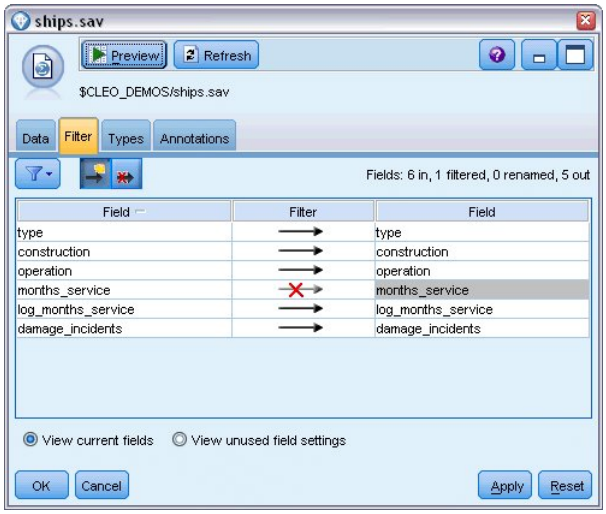


Figura 311. Filtrando um campo desnecessário

(Alternativamente, você poderia alterar a função para **Nenhum** para este campo na guia Tipos em vez de excluí-la, ou selecionar os campos que deseja utilizar no nó de modelagem.)

3. Na guia Tipos do nó de origem, configure a função para o campo *damage_incidentes* para **Destino**. Todos os outros campos devem ter seu papel configurado como **Entrada**.
4. Clique em **Valores de leitura** para instanciar os dados.

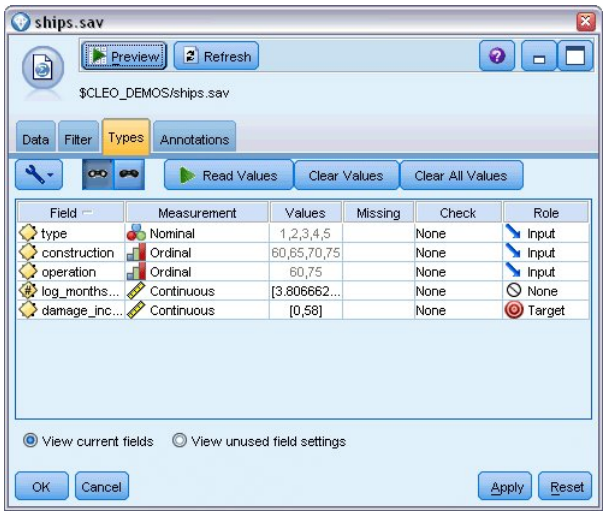


Figura 312. Configurando função de campo

5. Conecte um nó Genlin ao nó de origem; no nó Genlin, clique na guia **Modelo** .
6. Selecione *log_months_service* como a variável offset.

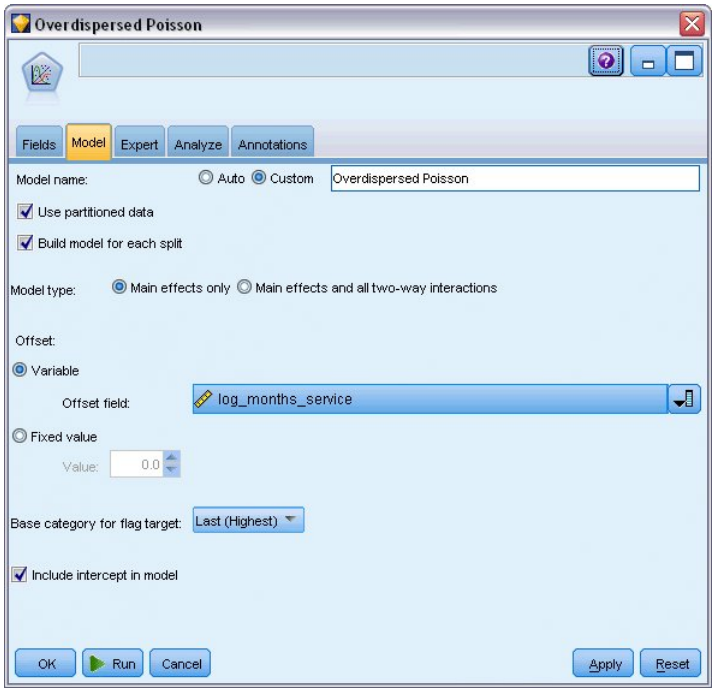


Figura 313. Escolhendo opções de modelo

7. Clique na guia **Expert** e selecione **Expert** para ativar as opções de modelagem de especialistas.

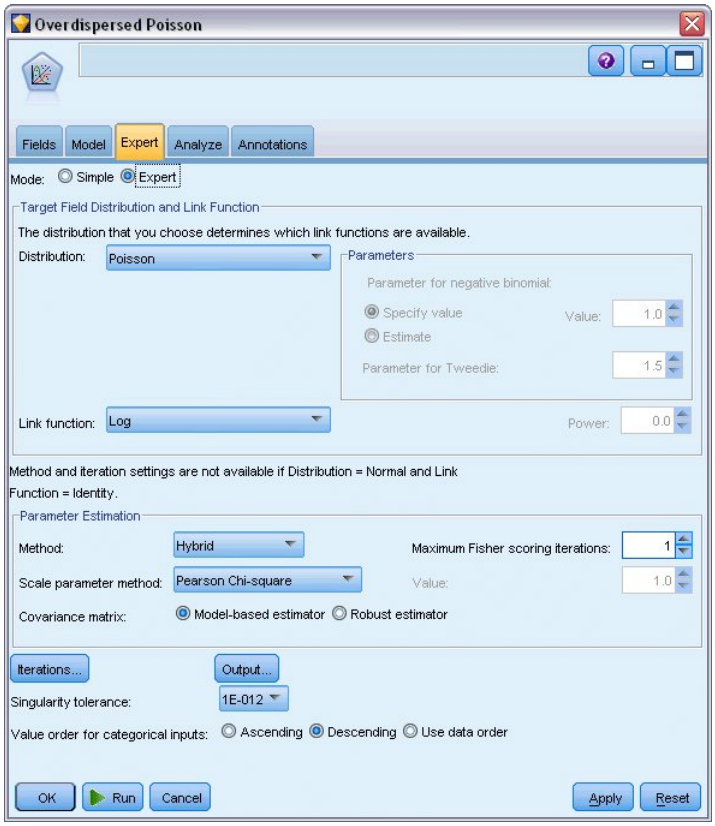


Figura 314. Escolhendo as opções de especialistas

8. Selecione **Poisson** como a distribuição para a resposta e **Log** como função de link.
9. Selecione **Pearson Chi-Square** como o método para estimar o parâmetro de escala. O parâmetro de escala geralmente é assumido como 1 em uma regressão de Poisson, mas McCullagh e Nelder usam a estimativa qui-quadrado de Pearson para obter estimativas de variância e níveis de significância mais conservadores.

10. Selecione **Descender** como a ordem de categoria para fatores. Isso indica que a primeira categoria de cada fator será a sua categoria de referência; o efeito dessa seleção sobre o modelo está na interpretação de estimativas de parâmetros.
11. Clique em **Executar** para criar o nugget do modelo, que é adicionado à tela do fluxo, e também à paleta de Modelos no canto superior direito. Para visualizar os detalhes do modelo, clique com o botão direito do mouse sobre o nugget e escolha **Editar** ou **Browse**, em seguida, clique na guia **Avançado** .

estatísticas de qualidade de ajuste

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	22.883	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	25.000	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Figura 315. estatísticas de qualidade de ajuste

A tabela de estatísticas de bondade-de-ajuste fornece medidas que são úteis para comparar modelos concorrentes. Adicionalmente, o *Value / df* para as estatísticas Deviance e Pearson Chi-Square dá estimativas correspondentes para o parâmetro scale. Esses valores devem estar próximos de 1.0 para uma regressão de Poisson; o fato de serem maiores que 1.0 indica que ajustar o modelo disperso em excesso pode ser razoável.

Teste de omnibus

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
63.650	8	.000

Dependent Variable: Number of damage incidents

Model: (Intercept), Year of construction, Period of operation, Ship type, offset = Logarithm of aggregate months of service

a. Compares the fitted model against the intercept-only model.

Figura 316. Teste de omnibus

O teste de onibus é um teste qui-quadrado de probabilidade do modelo atual versus o modelo nulo (neste caso, intercepto). O valor de significância menor que 0.05 indica que o modelo atual supera o modelo nulo.

Teste de efeitos do modelo

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	2138.657	1	.000
Year of construction	17.242	3	.001
Period of operation	6.249	1	.012
Ship type	15.415	4	.004

Dependent Variable: Number of damage incidents

Model: (Intercept), Year of construction, Period of operation, Ship type, offset = Logarithm of aggregate months of service

Figura 317. Teste de efeitos do modelo

Cada termo no modelo é testado para saber se ele tem algum efeito. Termos com valores de significado menores que 0.05 têm algum efeito discernível. Cada um dos termos de efeitos principais contribui para o modelo.

Estimativas de parâmetro

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-6.406	.2828	-6.960	-5.852	513.238	1	.000
[Year of construction=75]	.453	.3032	-.141	1.048	2.236	1	.135
[Year of construction=70]	.818	.2208	.386	1.251	13.743	1	.000
[Year of construction=65]	.697	.1946	.316	1.079	12.835	1	.000
[Year of construction=60]	0[a]	.	.	.	.	.	.
[Period of operation=75]	.384	.1538	.083	.686	6.249	1	.012
[Period of operation=60]	0[a]	.	.	.	.	.	.
[Ship type=5]	.326	.3067	-.276	.927	1.127	1	.288
[Ship type=4]	-.076	.3779	-.817	.665	.040	1	.841
[Ship type=3]	-.687	.4279	-1.526	.151	2.581	1	.108
[Ship type=2]	-.543	.2309	-.996	-.091	5.536	1	.019
[Ship type=1]	0[a]	.	.	.	.	.	.
(Scale)	1.691[b]						

Dependent Variable: Number of damage incidents

Model: (Intercept), Year of construction, Period of operation, Ship type, offset = Logarithm of aggregate months of service

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

Figura 318. Estimativas de parâmetro

A tabela de estimativas de parâmetros resume o efeito de cada preditor. Embora a interpretação dos coeficientes neste modelo seja difícil por causa da natureza da função de ligação, os sinais dos coeficientes para covariáveis e valores relativos dos coeficientes para os níveis de fatores podem dar insights importantes sobre os efeitos dos preditores no modelo.

- Para covariados, os coeficientes positivos (negativos) indicam relações positivas (inversa) entre os preditores e o resultado. Um valor crescente de um covariado com um coeficiente positivo corresponde a uma taxa crescente de incidentes de danos.
- Para os fatores, um nível de fator com um coeficiente maior indica maior incidência de danos. O sinal de um coeficiente para um nível de fator é dependente do efeito desse nível de fator relativo à categoria de referência.

Você pode fazer as seguintes interpretações com base nas estimativas do parâmetro:

- O tipo de envio *B [type=2]* tem uma taxa de dano estatisticamente significativa (*p* valor de 0.019) inferior (coeficiente estimado de-0.543) do tipo *A [type=1]*, a categoria de referência. O tipo *C [type=3]* realmente possui um parâmetro estimado menor que *B*, mas a variabilidade na estimativa de *C*obscurece o efeito. Veja os meios marginais estimados para todas as relações entre os níveis de fatores.
- Navios construídos entre 1965-69 *[construction=65]* e 1970-74 *[construction=70]* têm taxas de danos estatisticamente significativas (valores*p* <0.001) mais altas (coeficientes estimados de 0.697 e 0.818, respectivamente) do que aqueles construídos entre 1960-64 *[construction=60]*, a categoria de referência.... Veja os meios marginais estimados para todas as relações entre os níveis de fatores.
- Navios em operação entre 1975-79 *[operation=75]* têm taxas de danos estatisticamente significantes (valor*p* de 0.012) mais altas (coeficiente estimado de 0.384) do que aqueles em operação entre 1960-1974 *[operation=60]*

Modelos alternativos de encaixe

Um problema com a regressão de Poisson "superdisperso" é que não há uma maneira formal de testá-lo versus a regressão de Poisson "padrão". No entanto, um teste formal sugerido para determinar se há excesso de dispersão é realizar um teste de proporção de verossimilhança entre uma regressão Poisson "padrão" e uma regressão binomial negativa com todas as outras configurações iguais. Se não houver sobredispersão na regressão do Poisson, então a estatística – 2 × (log-verossimilhança para o modelo Poisson – log-verossimilhança para modelo binomial negativo) deve ter uma distribuição de mistura com metade de sua massa de probabilidade em 0 e o restante em uma distribuição qui-quadrado com 1 grau de liberdade.

1. Selecione **Valor Fixo** como o método para estimar o parâmetro de escala. Por padrão, esse valor é 1.

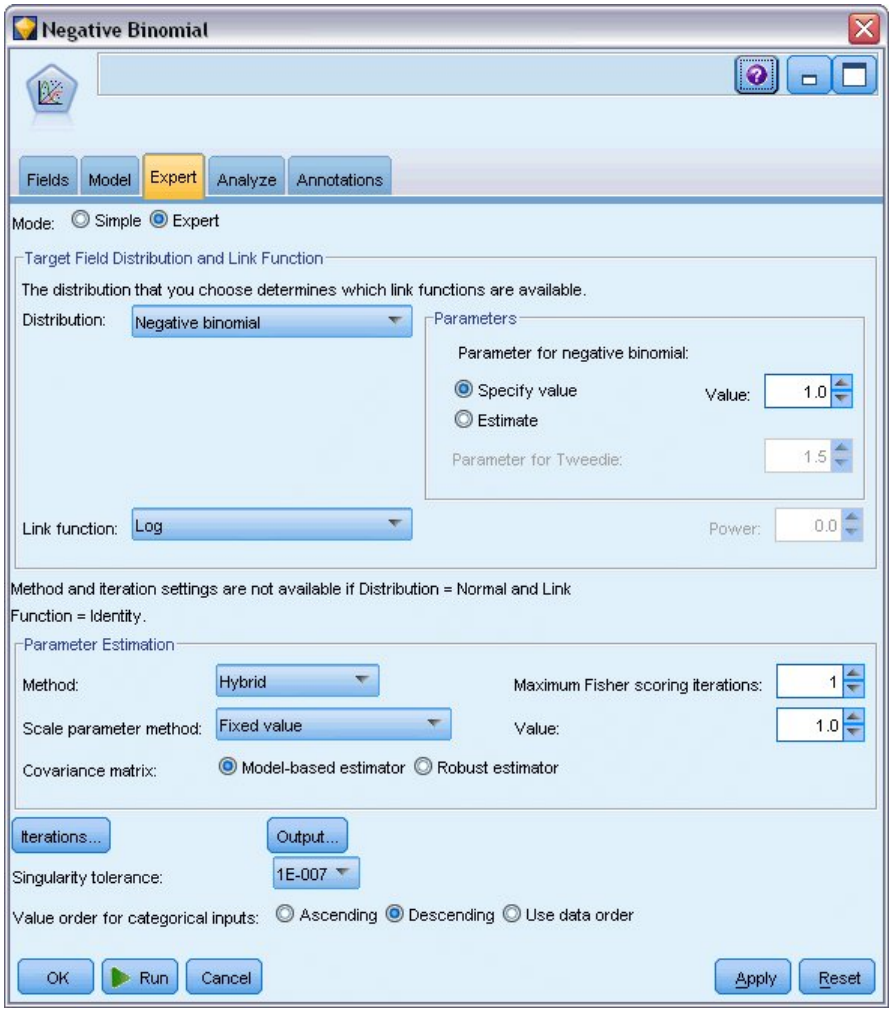


Figura 319. guia especialista

2. Para encaixar a regressão binomial negativa, copie e cole o nó Genlin, anexe-o ao nó de origem, abra o novo nó e clique na guia **Expert** .
3. Selecione **Binomial Negativo** como a distribuição. Deixe o valor padrão de 1 para o parâmetro ancillary.
4. Execute o fluxo e navegue na guia Avançado sobre os nuggets de modelo recém-criado.

estatísticas de qualidade de ajuste

	Value	df	Value/df
Deviance	38.695	25	1.548
Scaled Deviance	38.695	25	
Pearson Chi-Square	42.275	25	1.691
Scaled Pearson Chi-Square	42.275	25	
Log Likelihood[a]	-68.281		
Akaike's Information Criterion (AIC)	154.562		
Finite Sample Corrected AIC (AICC)	162.062		
Bayesian Information Criterion (BIC)	168.299		
Consistent AIC (CAIC)	177.299		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Figura 320. Estatísticas de bondade-de-ajuste para regressão de Poisson padrão

O log da verossimilhança relatado para a regressão Poisson padrão é-68.281. Compare isso com o modelo binomial negativo.

	Value	df	Value/df
Deviance	11.145	25	.446
Scaled Deviance	11.145	25	
Pearson Chi-Square	8.815	25	.353
Scaled Pearson Chi-Square	8.815	25	
Log Likelihood[a]	-83.725		
Akaike's Information Criterion (AIC)	185.450		
Finite Sample Corrected AIC (AICC)	192.950		
Bayesian Information Criterion (BIC)	199.187		
Consistent AIC (CAIC)	208.187		

Dependent Variable: Number of damage incidents
Model: (Intercept), type, construction, operation, offset = log_months_service

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Figura 321. Estatísticas de bondade-de-ajuste para regressão binomial negativa

O log da verossimilhança relatado para a regressão binomial negativa é-83.725 Esta é, na verdade, *menor* do que a log-probabilidade para a regressão Poisson, que indica (sem a necessidade de um teste de proporção de verossimilhança) que esta regressão binomial negativa não oferece uma melhoria sobre a regressão do Poisson.

No entanto, o valor escolhido de 1 para o parâmetro auxiliar da distribuição binomial negativa pode não ser o ideal para este dataset. Outra maneira que você poderia testar para a superdispersão é encaixar um modelo binomial negativo com parâmetro auxiliar igual a 0 e solicitar o teste de multiplicador de Lagrange na Diálogo de saída da guia Expert. Se o teste não for significativo, a sobredispersão não deve ser um problema para este dataset.

Resumo

Utilizando Modelos Lineares Generalizados, você se encaixou em três modelos diferentes para contagem de dados. A regressão binomial negativa foi mostrada para não oferecer qualquer melhoria sobre a regressão de Poisson. A regressão de Poisson superdispersa parece oferecer uma alternativa razoável para o modelo padrão Poisson, mas não há um teste formal para a escolha entre eles.

Explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no *IBM SPSS Modelador Algorithms Guide*.

Procedimentos Relacionados

O Procedimento Modelos Lineares Generalizados é uma ferramenta poderosa para encaixar uma variedade de modelos.

- O Procedimento de Equações de Estimativa Generalizadas amplia o modelo linear generalizado para permitir medições repetidas.
- O Procedimento de Modelos Mistas Lineares permite que você se encaixe em modelos para variáveis dependentes de escala com um componente aleatório e / ou medições repetidas.

Leituras recomendadas

Veja os textos a seguir para obter mais informações sobre modelos lineares generalizados:

Cameron, A. C., e P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2ª ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., e J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., e J. A. Nelder. 1989. *Generalized Linear Models*, 2ª ed. Londres: Chapman & Hall.

Capítulo 24. Encaixe uma regressão de Gamma para sinistros de seguros de carro (Modelos Lineares Generalizados)

Um modelo linear generalizado pode ser usado para se adequar a uma regressão Gamma para a análise de dados de alcance positivo. Por exemplo, um dataset apresentado e analisado em outro lugar[3]refere-se a pedidos de indenizações para automóveis. A quantidade de reclamação média pode ser modelada como tendo uma distribuição gama, utilizando uma função de link inversa para relacionar a média da variável dependente a uma combinação linear dos preditores. A fim de dar conta do número variado de reclamações usadas para computar os montantes de reclamação média, você especifica *Número de reclamações* como o peso escalonante.

Este exemplo usa o fluxo chamado *car-insurance_genlin.str*, que faz referência ao arquivo de dados denominado *car_insurance_claims.sav* O arquivo de dados está na pasta *Demos* e o arquivo stream está na subpasta *streams* .

Criando o Stream

1. Inclua um nó de origem do Arquivo de Estatísticas apontando para *car_insurance_claims.sav* na pasta *Demos* .

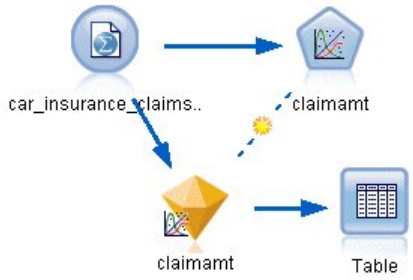


Figura 322. Fluxo de amostra para prever as reclamações de seguro de carro

2. Na guia Tipos do nó de origem, configure a função para o campo *claimamt* para **Target**. Todos os outros campos devem ter seu papel configurado como **Entrada**.
3. Clique em **Valores de leitura** para instanciar os dados.

[3] McCullagh, P., e J. A. Nelder. 1989. *Generalized Linear Models*, 2ª ed. Londres: Chapman & Hall.

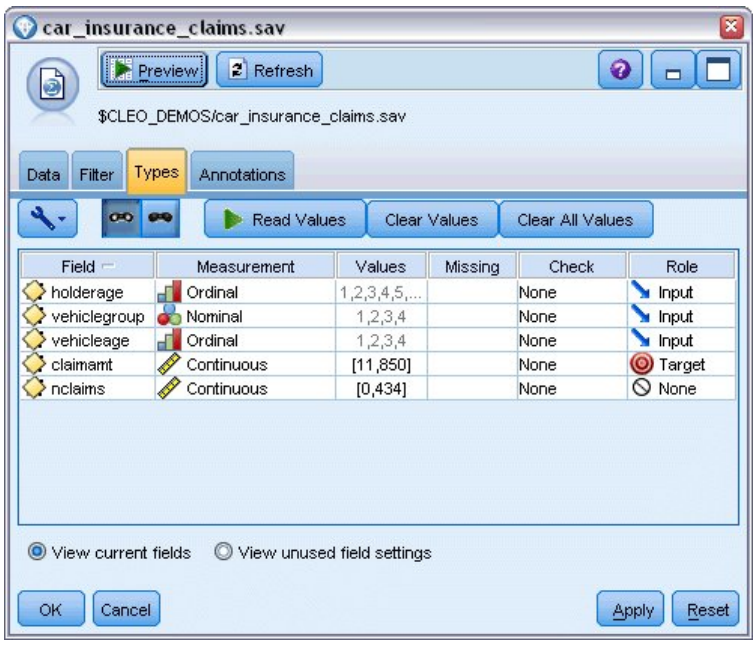


Figura 323. Configurando função de campo

4. Conecte um nó Genlin ao nó de origem; no nó Genlin, clique na guia Campos.
5. Selecione *nalega* como o campo de peso de escala.

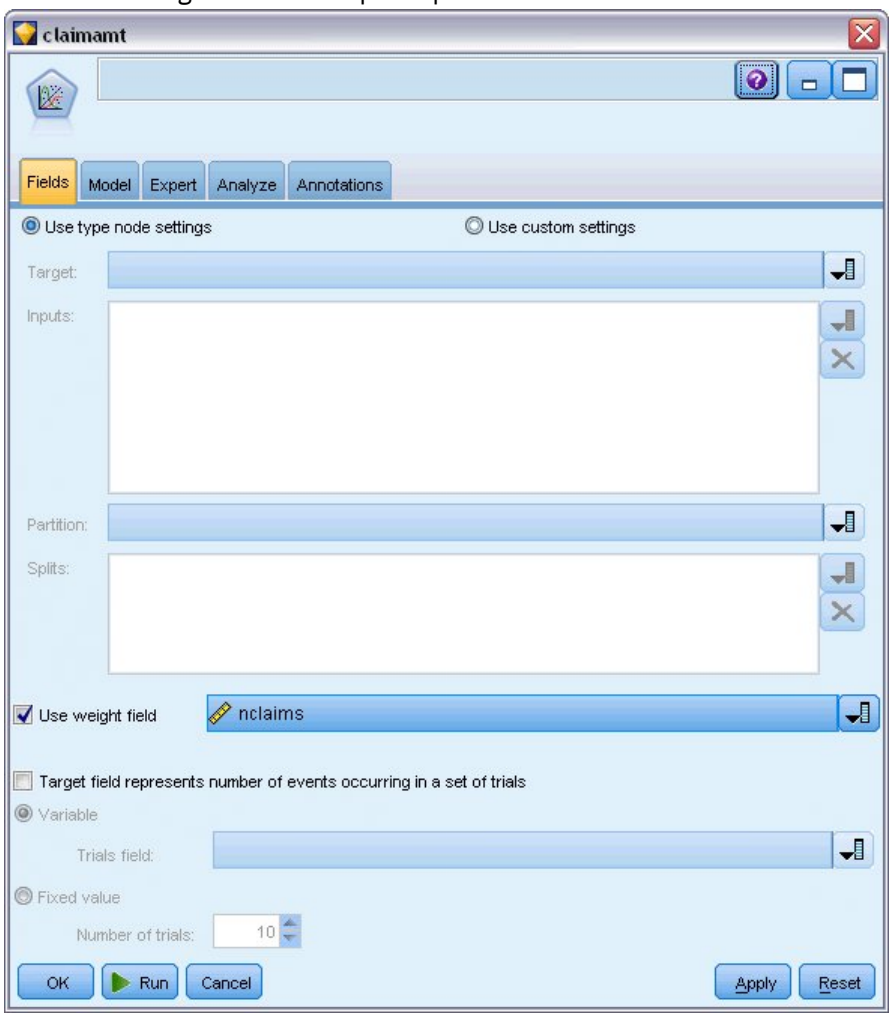


Figura 324. Escolhendo opções de campo

6. Clique na guia Expert e selecione **Expert** para ativar as opções de modelagem de especialistas.

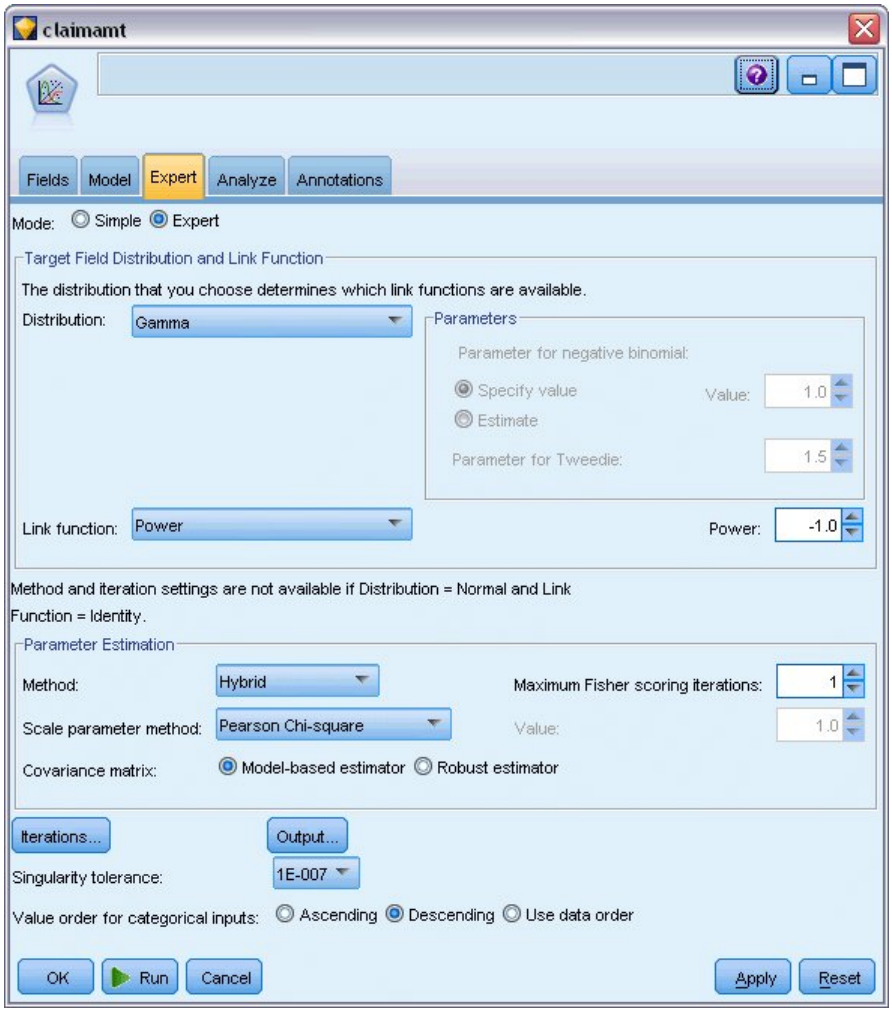


Figura 325. Escolhendo as opções de especialistas

7. Selecione **Gamma** como distribuição de resposta.
8. Selecione **Energia** como a função de ligação e digite `-1.0` como o expoente da função de energia. Este é um link inverso.
9. Selecione **Pearson chi-square** como o método para estimar o parâmetro de escala. Esse é o método usado por McCullagh e Nelder, portanto, seguimos aqui para replicar seus resultados.
10. Selecione **Descender** como a ordem de categoria para fatores. Isso indica que a primeira categoria de cada fator será a sua categoria de referência; o efeito dessa seleção sobre o modelo está na interpretação de estimativas de parâmetros.
11. Clique em **Executar** para criar o nugget modelo, que é adicionado à tela do fluxo, e também à paleta de Modelos no canto superior direito. Para visualizar os detalhes do modelo, clique com o botão direito do mouse sobre o nugget do modelo e escolha **Editar** ou **Procurar**, em seguida, selecione a guia Avançado.

Estimativas de parâmetro

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.003	.0004	.003	.004	66.593	1	.000
[Policyholder age=8]	.001	.0004	.000	.002	4.898	1	.027
[Policyholder age=7]	.001	.0004	.000	.002	5.046	1	.025
[Policyholder age=6]	.001	.0004	.000	.002	5.740	1	.017
[Policyholder age=5]	.001	.0004	.001	.002	10.682	1	.001
[Policyholder age=4]	.000	.0004	.000	.001	1.268	1	.260
[Policyholder age=3]	.000	.0004	.000	.001	.720	1	.396
[Policyholder age=2]	.000	.0004	-.001	.001	.054	1	.816
[Policyholder age=1]	0[a]	.	.	.	.	.	.
[Vehicle age=4]	.004	.0004	.003	.005	88.175	1	.000
[Vehicle age=3]	.002	.0002	.001	.002	53.013	1	.000
[Vehicle age=2]	.000	.0001	.000	.001	13.191	1	.000
[Vehicle age=1]	0[a]	.	.	.	.	.	.
[Vehicle group=4]	-.001	.0002	-.002	-.001	61.883	1	.000
[Vehicle group=3]	-.001	.0002	-.001	.000	13.039	1	.000
[Vehicle group=2]	3.765E-5	.0002	.000	.000	.050	1	.823
[Vehicle group=1]	0[a]	.	.	.	.	.	.
(Scale)	1.209[b]						

Dependent Variable: Average cost of claims

Model: (Intercept), Policyholder age, Vehicle age, Vehicle group

a. Set to zero because this parameter is redundant.

b. Computed based on the Pearson chi-square.

Figura 326. Estimativas de parâmetro

O teste de onibus e testes de efeitos de modelo (não mostrados) indicam que o modelo supera o modelo nulo e que cada um dos principais termos de efeitos contribuem para o modelo. A tabela de estimativas de parâmetros mostra os mesmos valores obtidos por McCullagh e Nelder para os níveis de fator e o parâmetro de escala.

Resumo

Usando Modelos Lineares Generalizados, você se encaixou em uma regressão de gama para os dados de reclamações. Observe que enquanto a função de link canônico para a distribuição gama foi usada neste modelo, um link de log também dará resultados razoáveis. Em geral, é difícil comparar diretamente modelos com diferentes funções de link; no entanto, o link de log é um caso especial do link de energia em que o expoente é 0, portanto, é possível comparar os desvios de um modelo com um link de log e um modelo com um link de energia para determinar qual fornece o melhor ajuste (veja, por exemplo, a seção 11.3 de McCullagh e Nelder).

Explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no *IBM SPSS Modelador Algorithms Guide*.

Procedimentos Relacionados

O Procedimento Modelos Lineares Generalizados é uma ferramenta poderosa para encaixar uma variedade de modelos.

- O Procedimento de Equações de Estimativa Generalizadas amplia o modelo linear generalizado para permitir medições repetidas.
- O Procedimento de Modelos Mistas Lineares permite que você se encaixe em modelos para variáveis dependentes de escala com um componente aleatório e / ou medições repetidas.

Leituras recomendadas

Veja os textos a seguir para obter mais informações sobre modelos lineares generalizados:

Cameron, A. C., e P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. Dobson, A. J. 2002. *An Introduction to Generalized Linear Models*, 2ª ed. Boca Raton, FL: Chapman & Hall/CRC. Hardin, J. W., e J. M. Hilbe. 2003. *Generalized Linear Models and Extension*. Station, TX: Stata Press. McCullagh, P., e J. A. Nelder. 1989. *Generalized Linear Models*, 2ª ed. Londres: Chapman & Hall.

Capítulo 25. Classificando Amostras De Células (SVM)

O Suporte Vector Machine (SVM) é uma técnica de classificação e regressão que é particularmente adequada para grandes conjuntos de dados. Um amplo conjunto de dados é um deles com um grande número de preditores, como podem ser encontrados no campo da bioinformática (a aplicação da tecnologia da informação aos dados bioquímicos e biológicos).

Um pesquisador médico obteve um conjunto de dados contendo características de um número de amostras de células humanas extraídas de pacientes que se acredita que estejam correndo risco de desenvolver câncer. A análise dos dados originais demonstrou que muitas das características diferiram significativamente entre amostras benignas e malignas. O pesquisador quer desenvolver um modelo SVM que possa utilizar os valores dessas características celulares em amostras de outros pacientes para dar uma indicação antecipada de se suas amostras podem ser benignas ou malignas.

Este exemplo usa o fluxo denominado *svm_cancer.str*, disponível na pasta *Demos* sob a subpasta *streams* O arquivo de dados é *cell_samples.data*.. Veja o tópico "Pasta Demos" na página 4 para obter mais informações.

O exemplo é baseado em um dataset que está disponível publicamente a partir do Repositório UCI Machine Learning . O dataset consiste em várias centenas de registros de amostras de células humanas, cada uma delas contém os valores de um conjunto de características celulares. Os campos em cada registro são:

Nome do campo	**Descrição**
ID	Identificador do paciente
Aglomerado	Espessura do clump
UnifSize	Uniformidade do tamanho da célula
UnifShape	Uniformidade da forma celular
MargAdh	Aderência marginal
SingEpiSize	Tamanho único do celular epitelial
BareNuc	Núcleos nus
BlandChrom	Cromatina de bland
NormNucl	Nucleoli normal
MIT	Mitoses
Classe	Benigno ou maligno

Para os propósitos deste exemplo, estamos usando um dataset que tem um número relativamente pequeno de preditores em cada registro.

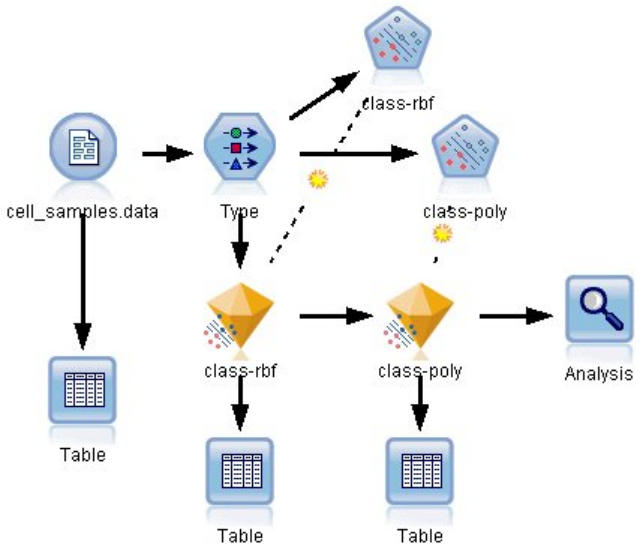
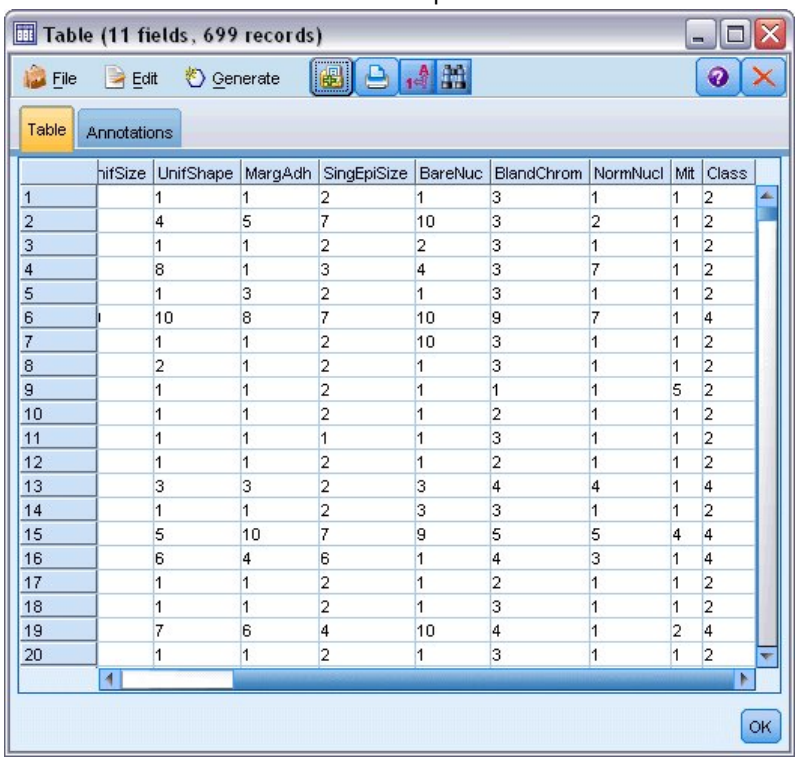


Figura 327. Fluxo de amostra para mostrar a modelagem SVM

1. Crie um novo fluxo e inclua um nó de origem do Arquivo Var apontando para *cell_samples.data* na pasta *Demos* de sua instalação do IBM SPSS Modelador .

   Vamos dar uma olhada nos dados do arquivo de origem.

2. Adicionar um nó da Tabela no fluxo.
3. Conecte o nó da Tabela ao nó do Arquivo Var e execute o fluxo.



	hifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1		1	1	2	1	3	1	1	2
2		4	5	7	10	3	2	1	2
3		1	1	2	2	3	1	1	2
4		8	1	3	4	3	7	1	2
5		1	3	2	1	3	1	1	2
6		10	8	7	10	9	7	1	4
7		1	1	2	10	3	1	1	2
8		2	1	2	1	3	1	1	2
9		1	1	2	1	1	1	5	2
10		1	1	2	1	2	1	1	2
11		1	1	1	1	3	1	1	2
12		1	1	2	1	2	1	1	2
13		3	3	2	3	4	4	1	4
14		1	1	2	3	3	1	1	2
15		5	10	7	9	5	5	4	4
16		6	4	6	1	4	3	1	4
17		1	1	2	1	2	1	1	2
18		1	1	2	1	3	1	1	2
19		7	6	4	10	4	1	2	4
20		1	1	2	1	3	1	1	2

Figura 328. Dados de origem para SVM

O campo *ID* contém os identificadores de pacientes. As características das amostras de células de cada paciente estão contidas nos campos *Clump* a *Mit*. Os valores são gradados de 1 10, sendo 1 os mais próximos de benignos.

O campo *Classe* contém o diagnóstico, como confirmado por procedimentos médicos separados, quanto a se as amostras são benignas (valor = 2) ou malignas (valor = 4).

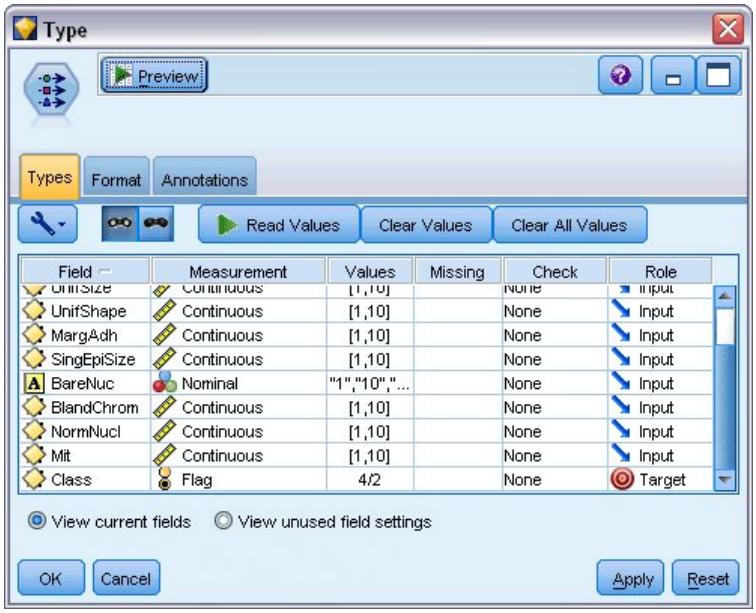


Figura 329. Configurações do nó do tipo

4. Adicire um nó Tipo e anexe-o no nó do Arquivo Var.
5. Abra o nó Tipo.

   Queremos que o modelo preveja o valor de *Classe* (isto é, benigno (= 2) ou maligno (= 4)). Como esse campo pode ter um dos dois únicos valores possíveis, precisamos mudar o seu nível de medição para refletir isso.
6. Na coluna **Measurement** para o campo *Class* (o último na lista), clique no valor **Continuous** e altere-o para **Bandeira**.
7. Clique em **Ler valores**.
8. Na coluna **Função** , configure a função para *ID* (o identificador do paciente) para **Nenhum**, uma vez que este não será usado tanto como um preditor ou um destino para o modelo.
9. Configure a função para o destino, *Classe*, para **Alvo** e deixe o papel de todos os outros campos (os preditores) como **Entrada**.
10. Clique em **OK**.

    O nó SVM oferece uma escolha de funções do kernel para a realização de seu processamento. Como não há uma maneira fácil de saber qual função desempenha melhor com qualquer dado conjunto de dados, escolheremos funções diferentes por sua vez e compararemos os resultados. Vamos começar com a inadimplência, RBF (Função Radial Base).

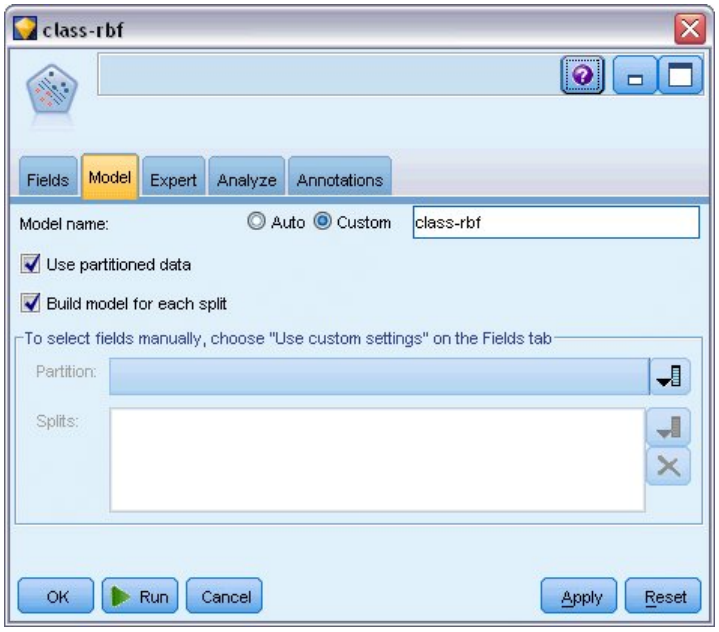


Figura 330. Configurações da guia do modelo

11. A partir da paleta Modelagem, anexe um nó SVM ao nó Type.
12. Abra o nó SVM. Na guia **Modelo** , clique na opção **Custom** para **Nome do modelo** e digite *class-rbf* no campo de texto adjacente.

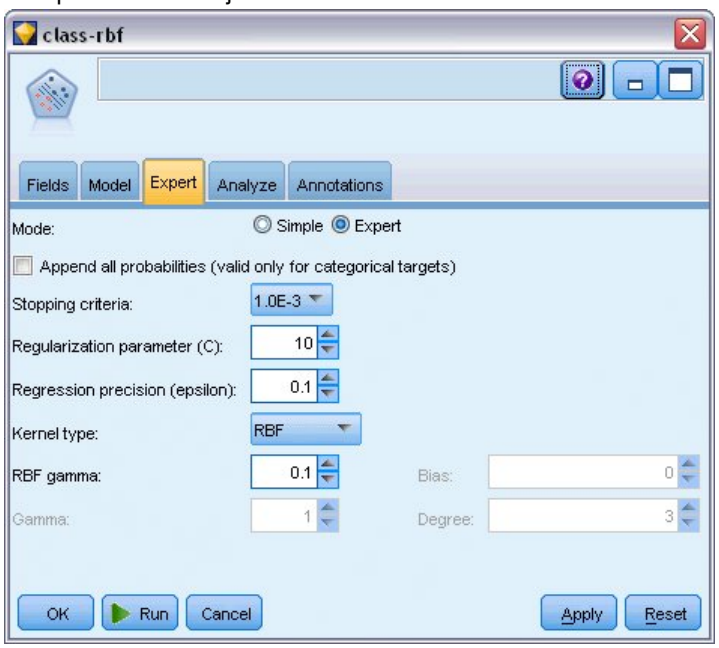


Figura 331. Configurações da guia Expert Padrão

13. Na guia **Expert** , configure o **Mode** para **Expert** para a legibilidade mas deixe todas as opções padrão como elas são. Note que **Kernel type** é configurado como **RBF** por padrão. Todas as opções estão esmaecidas no modo Simples.

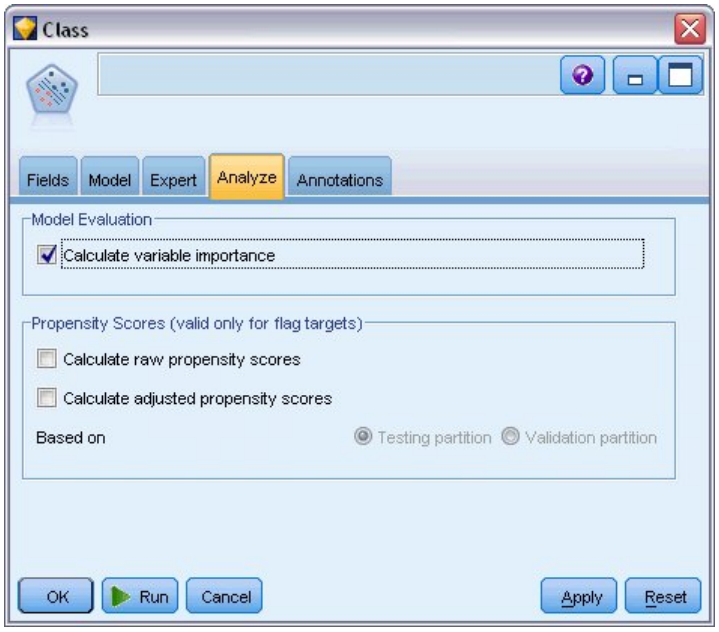


Figura 332. Analisar configurações da guia

14. Na guia **Analisar** , selecione a caixa de seleção **Calcule importância variável** .
15. Clique em **Executar** . O nugget modelo é colocado no fluxo, e na paleta de Models na parte superior direita da tela.
16. Clique duas vezes no nugget modelo no fluxo.

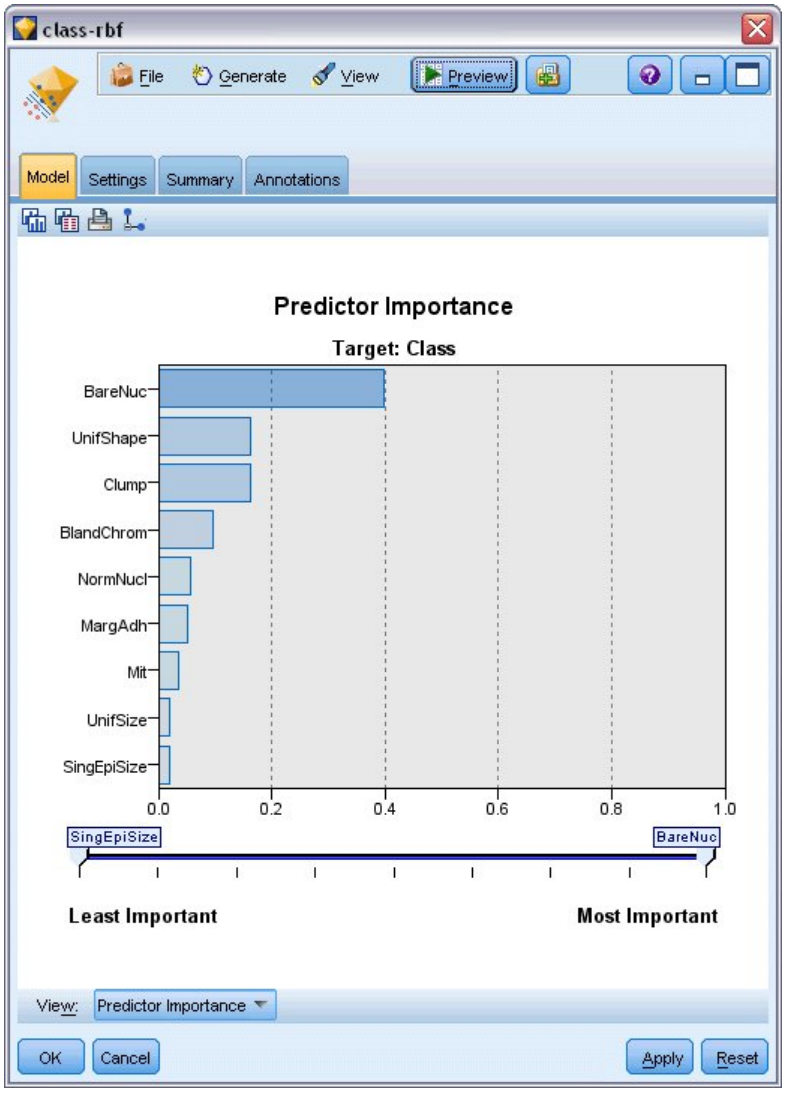
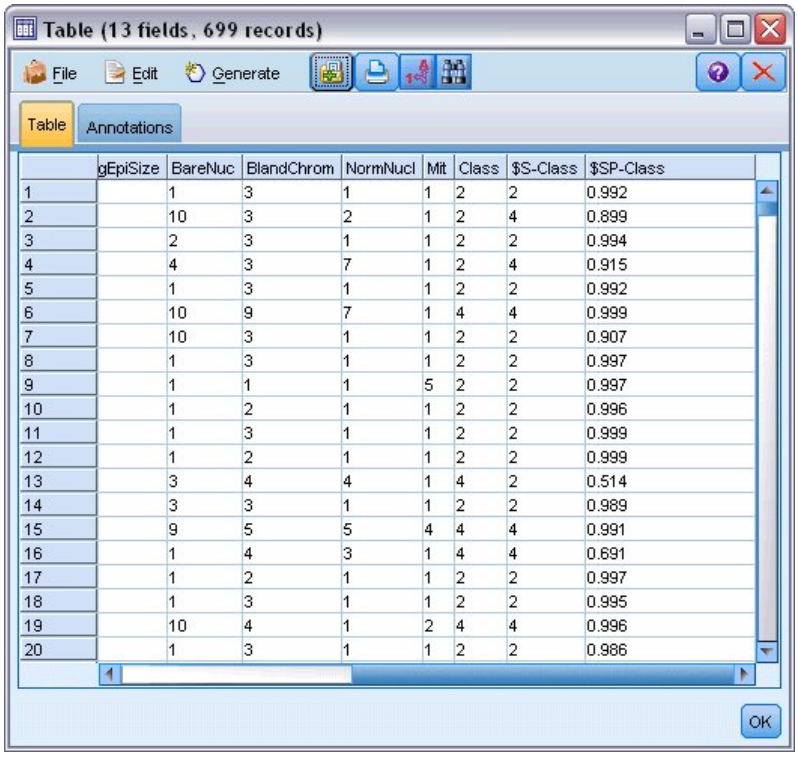


Figura 333. Gráfico Importância do Previdor

Na guia Modelo, o gráfico de Importância Previdente mostra o efeito relativo dos vários campos sobre a previsão. Isso nos mostra que *BareNuc* tem facilmente o maior efeito, enquanto *UnifShape* e *Clump* também são bastante significativos

1. Clique em **OK**.
2. Conecte um nó da Tabela na nugget de modelo *class-rbf* .
3. Abra o nó da Tabela e clique em **Executar**.



	gEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class	$S-Class	$SP-Class
1		1	3	1	1	2	2	0.992
2		10	3	2	1	2	4	0.899
3		2	3	1	1	2	2	0.994
4		4	3	7	1	2	4	0.915
5		1	3	1	1	2	2	0.992
6		10	9	7	1	4	4	0.999
7		10	3	1	1	2	2	0.907
8		1	3	1	1	2	2	0.997
9		1	1	1	5	2	2	0.997
10		1	2	1	1	2	2	0.996
11		1	3	1	1	2	2	0.999
12		1	2	1	1	2	2	0.999
13		3	4	4	1	4	2	0.514
14		3	3	1	1	2	2	0.989
15		9	5	5	4	4	4	0.991
16		1	4	3	1	4	4	0.691
17		1	2	1	1	2	2	0.997
18		1	3	1	1	2	2	0.995
19		10	4	1	2	4	4	0.996
20		1	3	1	1	2	2	0.986

Figura 334. Campos adicionados para predição e valor de confiança

4. O modelo criou dois campos extras. Role a saída da tabela para o direito de vê-los:

Novo nome de campo	**Descrição**
classe $S	Valor para *Classe* previsto pelo modelo.
$SP-Classe	Escore de propensão para essa predição (a probabilidade dessa predição ser verdadeira, um valor de 0.0 a 1.0).

Apenas olhando para a tabela, podemos ver que as pontuações de propensão (na coluna *$SP-Class*) para a maioria dos registros são razoavelmente altas.

No entanto, há algumas exceções significativas; por exemplo, o registro para o paciente 1041801 na linha 13, em que o valor de 0.514 é inaceitavelmente baixo. Também, comparando *Classe* com *$S-Class*, é claro que este modelo fez uma série de previsões incorretas, mesmo onde a pontuação de propensão foi relativamente alta (por exemplo, linhas 2 e 4).

Vamos ver se conseguimos fazer melhor escolhendo um tipo de função diferente.

Tentando uma Função diferente

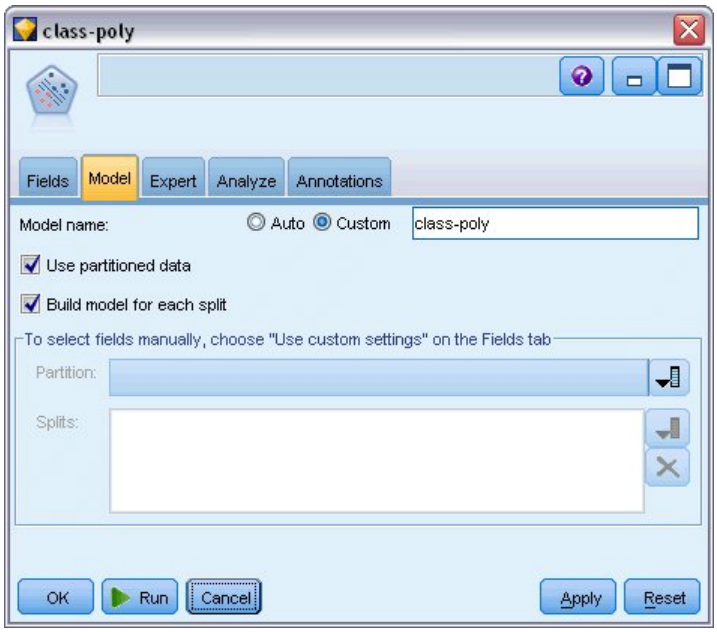


Figura 335. Como configurar um novo nome para o modelo

1. Fechar a janela de saída da Tabela.
2. Anexar um segundo nó de modelagem SVM ao nó Type.
3. Abra o novo nó SVM.
4. Na guia **Modelo** , escolha Custom e digite *class-poly* como o nome do modelo.

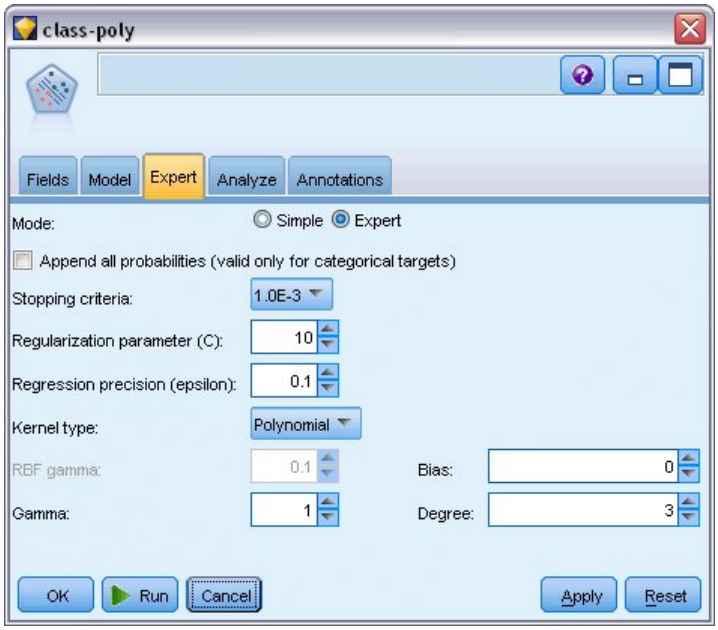
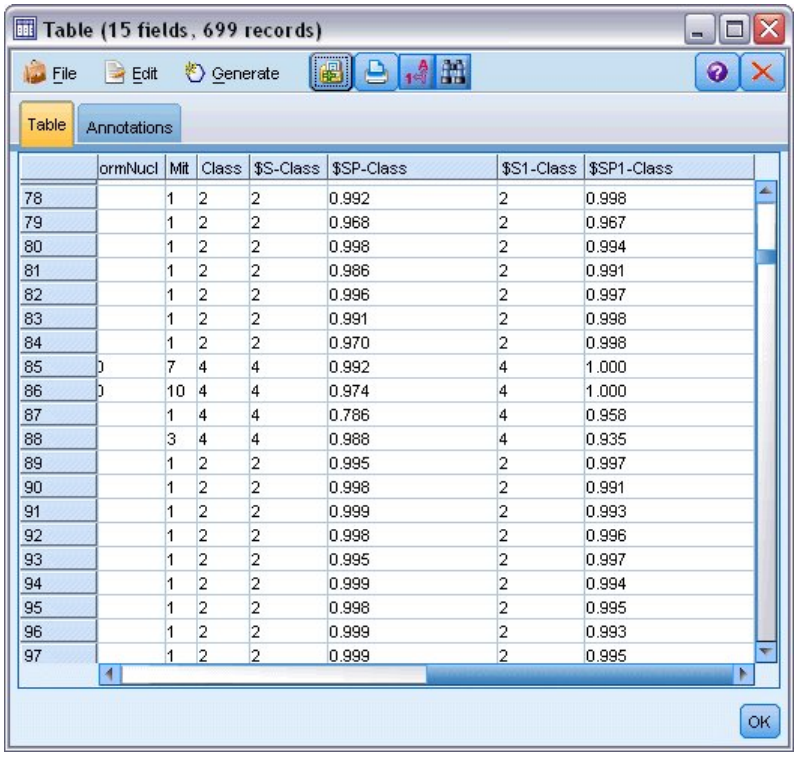


Figura 336. Configurações de guia de especialistas para Polinomial

5. Na guia **Expert** , configure **Mode** para **Expert**.
6. Configure **Tipo de Kernel** para **Polynomial** e clique em **Executar**. O nugget de modelo *class-poly* é adicionado ao fluxo e também à paleta de Modelos na parte superior direita da tela.
7. Conecte o nugget do modelo *class-rbf* ao nugget de modelo *class-poly* (escolha **Substituir** no diálogo de aviso).
8. Conecte um nó da Tabela na nugget de *classe-poly* .

9. Abra o nó da Tabela e clique em **Executar**.

Comparando os Resultados



	ormNucl	Mit	Class	$S-Class	$SP-Class	$S1-Class	$SP1-Class
78		1	2	2	0.992	2	0.998
79		1	2	2	0.968	2	0.967
80		1	2	2	0.998	2	0.994
81		1	2	2	0.986	2	0.991
82		1	2	2	0.996	2	0.997
83		1	2	2	0.991	2	0.998
84		1	2	2	0.970	2	0.998
85	)	7	4	4	0.992	4	1.000
86	)	10	4	4	0.974	4	1.000
87		1	4	4	0.786	4	0.958
88		3	4	4	0.988	4	0.935
89		1	2	2	0.995	2	0.997
90		1	2	2	0.998	2	0.991
91		1	2	2	0.999	2	0.993
92		1	2	2	0.998	2	0.996
93		1	2	2	0.995	2	0.997
94		1	2	2	0.999	2	0.994
95		1	2	2	0.998	2	0.995
96		1	2	2	0.999	2	0.993
97		1	2	2	0.999	2	0.995

Figura 337. Campos adicionados para função Polinomial

1. Role a saída da tabela para o direito de ver os campos recém-adicionados.

   Os campos gerados para o tipo de função Polynomial são nomeados *$S1-Class* e *$SP1-Class*.

   Os resultados para a Polinomial parecem muito melhores. Muitos dos escores de propensão são 0.995 ou melhor, o que é muito encorajante
2. Para confirmar a melhoria no modelo, anexe um nó de Análise na nugget do modelo *class-poly* .

Abra o nó da Análise e clique em **Executar**.

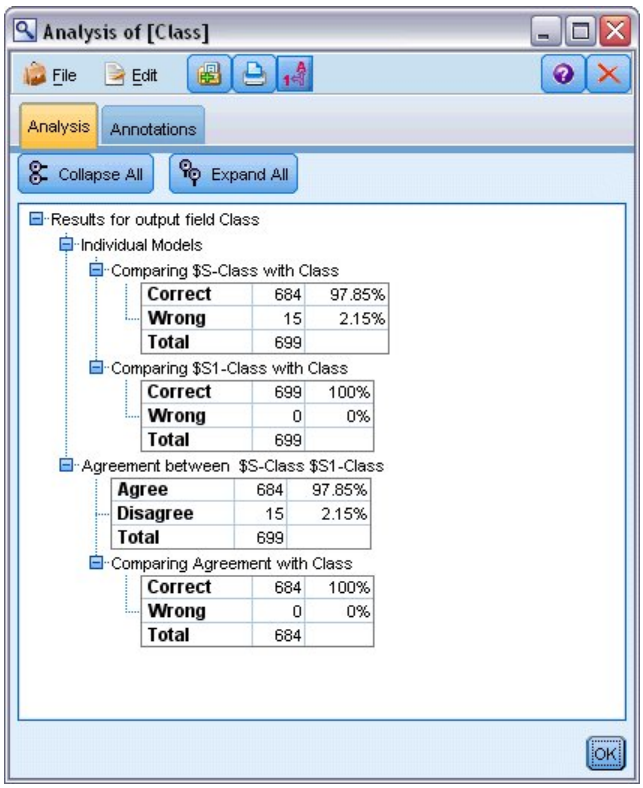


Figura 338. Nó de Análise

Essa técnica com o nó Análise possibilita comparar dois ou mais nuggets de modelo do mesmo tipo. A saída do nó Análise mostra que a função RBF prevê corretamente 97.85% dos casos, o que ainda é bastante bom. No entanto, a saída mostra que a função Polinomial previu corretamente o diagnóstico em cada caso único. Na prática você é improvável de ver 100% de precisão, mas você pode usar o nó Análise para ajudar a determinar se o modelo é aceitavelmente preciso para a sua aplicação particular.

Na verdade, nenhum dos outros tipos de função (Sigmoid e Linear) executa tão bem como Polynomial neste determinado dataset. No entanto, com um dataset diferente, os resultados poderiam facilmente ser diferentes, por isso sempre vale a pena tentar a gama completa de opções.

Resumo

Você tem usado diferentes tipos de funções do kernel SVM para prever uma classificação a partir de vários atributos. Você viu como diferentes kernels dão resultados diferentes para o mesmo dataset e como você pode medir a melhoria de um modelo sobre outro.

Capítulo 26. Usando a Cox Regression para modelar o tempo de rotatividade do cliente

Como parte de seus esforços para reduzir a migração para o concorrente, uma empresa de telecomunicações deseja modelar o "tempo para migração para o concorrente" para determinar os fatores que estão associados aos clientes que mais rapidamente querem mudar para outro serviço. Para isso, uma amostra aleatória de clientes é selecionada e seu tempo gasto como clientes, se ainda são clientes ativos, e vários outros campos são retirados do banco de dados.

Este exemplo usa o fluxo *telco_coxreg.str*, que faz referência ao arquivo de dados *telco.sav*.. O arquivo de dados está na pasta *Demos* e o arquivo stream está na subpasta *streams* . Veja o tópico "Pasta Demos" na página 4 para obter mais informações.

Construindo um Modelo Adequado

1. Inclua um nó de origem do Arquivo de Estatísticas apontando para *telco.sav* na pasta *Demos* .

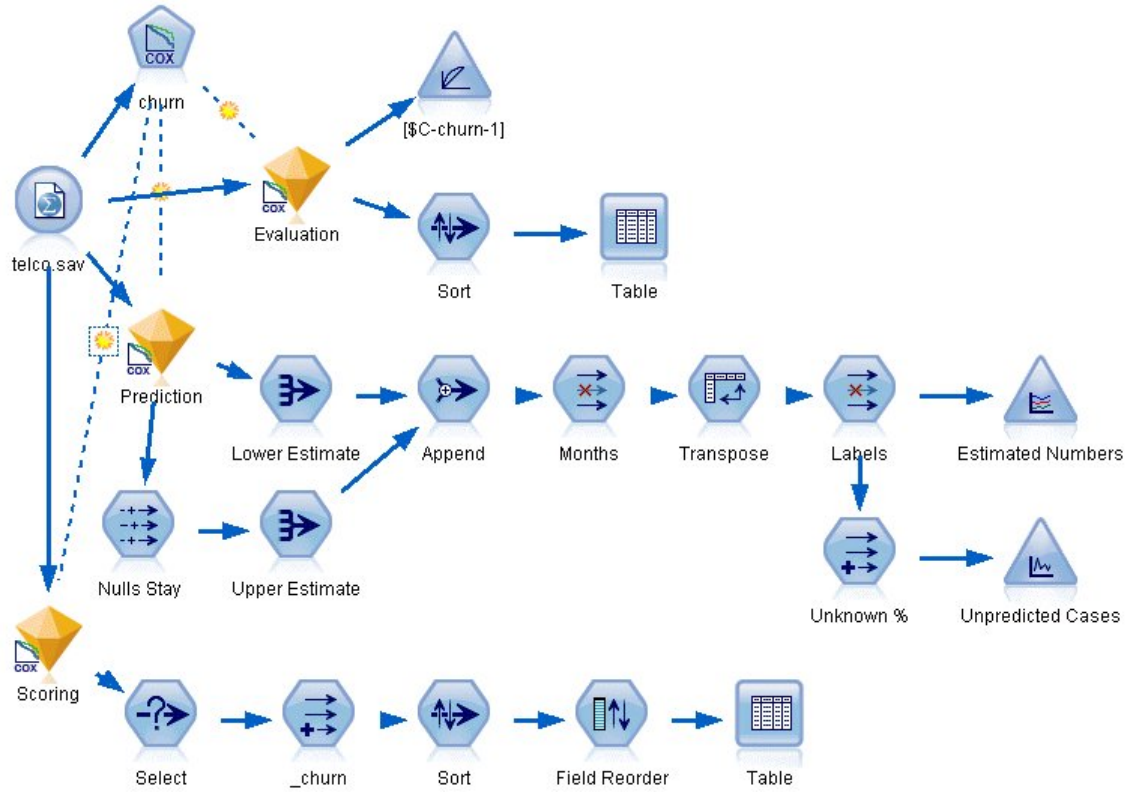


Figura 339. Fluxo de amostra para analisar tempo para churn

2. Na guia Filtro do nó de origem, exclua os campos *região*, *renda*, *longten* através do *wireten*, e *loglong* através de *logwire*.

Figura 340. Filtrando campos desnecessários

(Alternativamente, você poderia alterar a função para **Nenhum** para esses campos na guia Tipos em vez de excluí-la, ou selecionar os campos que deseja utilizar no nó de modelagem.)

3. Na guia Tipos do nó de origem, configure a função para o campo *churn* para **Target** e configure seu nível de medição para **Flag**. Todos os outros campos devem ter seu papel configurado como **Entrada**.
4. Clique em **Valores de leitura** para instanciar os dados.

Figura 341. Configurando função de campo

5. Anexe um nó Cox ao nó de origem; na guia **Fields (Campos**), selecione *tenure (resistência)* como a variável de tempo de sobrevivência.

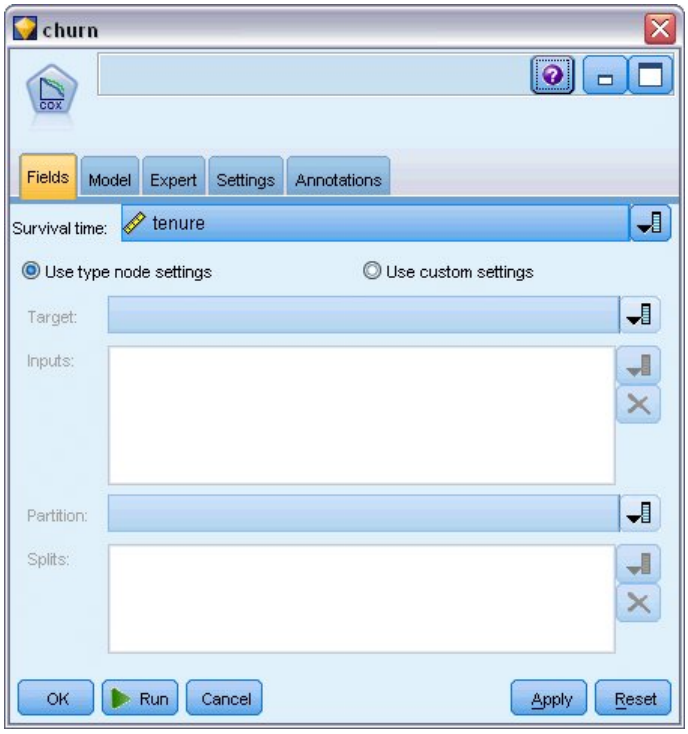


Figura 342. Escolhendo opções de campo

6. Clique na guia **Modelo**.
7. Selecione **Stepwise** como o método de seleção de variáveis.

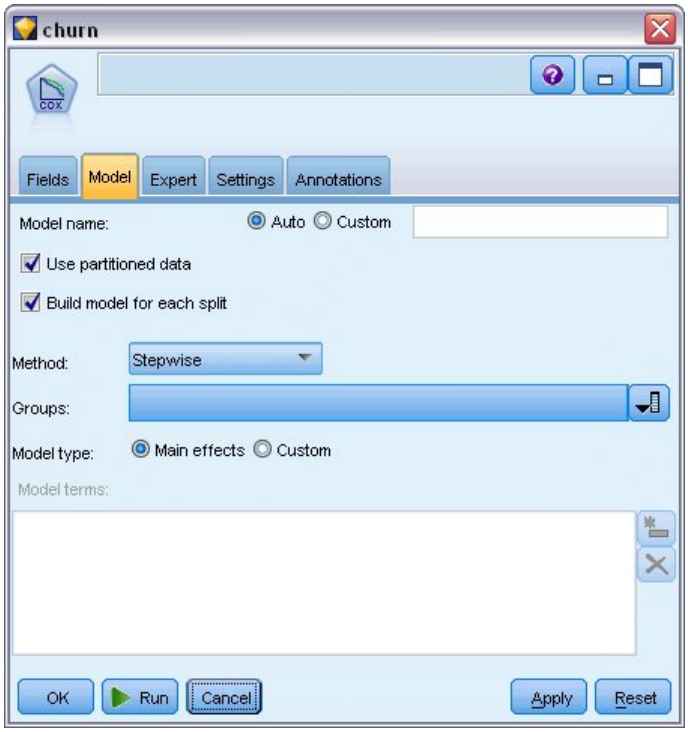


Figura 343. Escolhendo opções de modelo

8. Clique na guia **Expert** e selecione **Expert** para ativar as opções de modelagem de especialistas.
9. Clique em **Saída** .

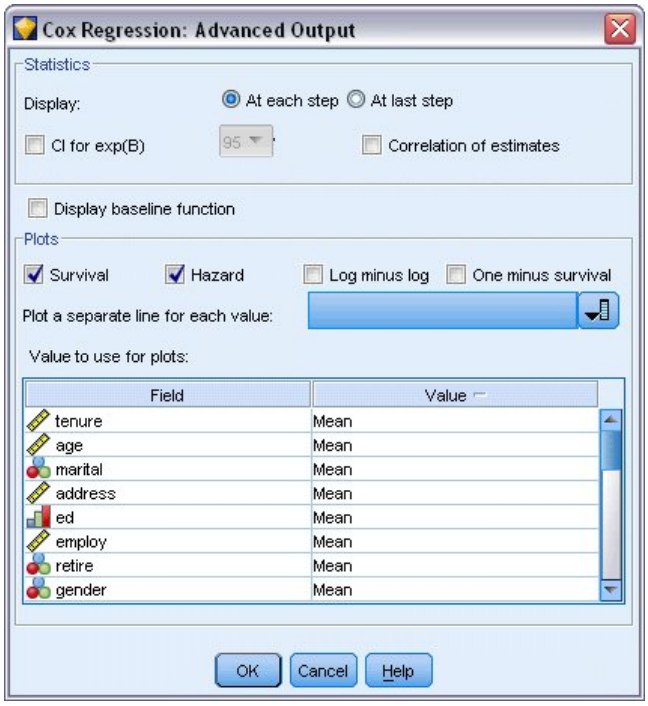


Figura 344. Escolhendo opções avançadas de saída

10. Selecione **Survival** e **Hazard** como tramas para produzir, em seguida, clique em **OK**.
11. Clique em **Executar** para criar o nugget modelo, que é adicionado ao fluxo, e para a paleta de Modelos no canto superior direito. Para visualizar seus detalhes, dê um duplo clique sobre o nugget no fluxo. Primeiro, veja a guia de saída Avançada.

casos censurados

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	274	27.4%
	Censored	726	72.6%
	Total	1000	100.0%
Cases dropped	Cases with missing values	0	0.0%
	Cases with negative time	0	0.0%
	Censored cases before the earliest event in a stratum	0	0.0%
	Total	0	0.0%
Total		1000	100.0%

a. Dependent Variable: Months with service

Figura 345. Sumarização do processamento de caso

A variável de status identifica se o evento ocorreu para um determinado caso. Caso o evento não tenha ocorrido, o caso é dito para ser censurado. Os casos censurados não são usados no cálculo dos coeficientes de regressão mas são usados para computar o risco de linha de base. O resumo do processamento de casos mostra que 726 casos são censurados. São clientes que não se chureceram.

Codificações de variáveis categóricas

		Frequency	(1)[b]	(2)	(3)	(4)
marital[a]	0=Unmarried	505	1			
	1=Married	495	0			
ed[a]	1=Did not complete high school	204	1	0	0	0
	2=High school degree	287	0	1	0	0
	3=Some college	209	0	0	1	0
	4=College degree	234	0	0	0	1
	5=Post-undergraduate degree	66	0	0	0	0
retire[a]	.00=No	953	1			
	1.00=Yes	47	0			
gender[a]	0=Male	483	1			
	1=Female	517	0			
tollfree[a]	0=No	526	1			
	1=Yes	474	0			
equip[a]	0=No	614	1			
	1=Yes	386	0			
callcard[a]	0=No	322	1			
	1=Yes	678	0			
wireless[a]	0=No	704	1			
	1=Yes	296	0			
multline[a]	0=No	525	1			
	1=Yes	475	0			
voice[a]	0=No	696	1			
	1=Yes	304	0			
pager[a]	0=No	739	1			
	1=Yes	261	0			
internet[a]	0=No	632	1			
	1=Yes	368	0			
callid[a]	0=No	519	1			
	1=Yes	481	0			
callwait[a]	0=No	515	1			
	1=Yes	485	0			
forward[a]	0=No	507	1			
	1=Yes	493	0			
confer[a]	0=No	498	1			
	1=Yes	502	0			
ebill[a]	0=No	629	1			
	1=Yes	371	0			
custcat[a]	1=Basic service	266	1	0	0	
	2=E-service	217	0	1	0	
	3=Plus service	281	0	0	1	
	4=Total service	236	0	0	0	

Figura 346. Codificações de variável categórica

As codinhas de variáveis categóricas são uma referência útil para interpretar os coeficientes de regressão para covariados categóricos, particularmente variáveis dicotomias. Por padrão, a categoria de referência é a categoria "última". Assim, por exemplo, mesmo que os clientes *Casados* tenham valores variáveis de 1 no arquivo de dados, eles são codificados como 0 para os propósitos da regressão.

seleção de variáveis

Step	-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
		Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1[a]	3392.536	162.303	1	.000	133.828	1	.000	133.828	1	.000
2[b]	3087.314	249.392	2	.000	305.222	1	.000	439.050	2	.000
3[c]	3027.085	328.426	3	.000	60.229	1	.000	499.279	3	.000
4[d]	2990.790	347.197	4	.000	36.294	1	.000	535.574	4	.000
5[e]	2973.790	362.673	5	.000	17.000	1	.000	552.574	5	.000
6[f]	2958.796	376.140	6	.000	14.994	1	.000	567.568	6	.000
7[g]	2945.503	384.717	7	.000	13.293	1	.000	580.861	7	.000
8[h]	2936.993	417.341	8	.000	8.510	1	.004	589.371	8	.000
9[i]	2926.000	423.911	9	.000	10.994	1	.001	600.364	9	.000
10[j]	2917.551	428.078	10	.000	8.449	1	.004	608.813	10	.000
11[k]	2913.308	436.837	11	.000	4.243	1	.039	613.056	11	.000
12[l]	2908.078	440.158	12	.000	5.230	1	.022	618.286	12	.000

a. Variable(s) Entered at Step Number 1: callcard
b. Variable(s) Entered at Step Number 2: longmon
c. Variable(s) Entered at Step Number 3: equip
d. Variable(s) Entered at Step Number 4: employ
e. Variable(s) Entered at Step Number 5: multline
f. Variable(s) Entered at Step Number 6: voice
g. Variable(s) Entered at Step Number 7: address
h. Variable(s) Entered at Step Number 8: equipmon
i. Variable(s) Entered at Step Number 9: ebill
j. Variable(s) Entered at Step Number 10: callid
k. Variable(s) Entered at Step Number 11: internet
l. Variable(s) Entered at Step Number 12: reside
m. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 3526.364
n. Beginning Block Number 1. Method = Forward Stepwise (Likelihood Ratio)

Figura 347. Testes de onibus

O processo de construção de modelo emprega um algoritmo stepwise adiante. Os testes onibus são medidas de quão bem o modelo executa. A mudança qui-quadrada da etapa anterior é a diferença entre o – 2 log-verossimilhança do modelo na etapa anterior e a etapa atual. Se a etapa foi incluir uma variável, a inclusão fará sentido se a significância da mudança for menor que 0.05. Se a etapa fosse remover uma variável, a exclusão faria sentido se a significância da mudança fosse maior que 0.10. Em doze etapas, doze variáveis são adicionadas ao modelo.

		B	SE	Wald	df	Sig.	Exp(B)
Step 12	address	-.035	.009	14.543	1	.000	.966
	employ	-.051	.010	25.767	1	.000	.950
	reside	-.103	.046	5.037	1	.025	.902
	equip	-1.948	.381	26.180	1	.000	.143
	callcard	.777	.151	26.451	1	.000	2.175
	longmon	-.233	.022	115.619	1	.000	.792
	equipmon	-.042	.011	15.377	1	.000	.959
	multline	.612	.145	17.854	1	.000	1.844
	voice	-.501	.157	10.197	1	.001	.606
	internet	-.362	.160	5.114	1	.024	.697
	callid	-.464	.148	9.790	1	.002	.629
	ebill	-.399	.156	6.557	1	.010	.671

Figura 348. Variáveis na equação (etapa 12 apenas)

O modelo final inclui *endereço*, *empregar*, *residem*, *equip*, *callcard*, *longmon*, *equipmon*, *multline*, *voice*, *internet*, *callid*, e *ebill*. Para entender os efeitos de preditores individuais, veja a Exp (B), que pode ser interpretada como a mudança prevista no risco para um aumento unitário do preditor.

- O valor de Exp (B) para *address* significa que o risco de perda de clientes é reduzido em 100%-(100% ×0.966) =3.4% para cada ano em que um cliente viveu no mesmo endereço. O risco de perda de clientes que viveram no mesmo endereço por cinco anos é reduzido em 100%-(100% $\times 0.966^5$) =15.88%.
- O valor de Exp (B) para *cartão de chamada* significa que o risco de perda de clientes para um cliente que não assina o serviço de cartão de chamada é 2.175 vezes o de um cliente com o serviço. Recheio das codinções de variáveis categóricas que *Não* = 1 para a regressão.

- O valor de Exp (B) para *internet* significa que o risco de perda de clientes para um cliente que não assina o serviço de Internet é 0.697 vezes o de um cliente com o serviço. Isso é um tanto preocupante porque sugere que os clientes com o serviço estão deixando a empresa mais rápido do que os clientes sem o serviço.

		Score	df	Sig.
Step 12	age	.122	1	.726
	marital	.648	1	.421
	income	1.476	1	.224
	ed	6.328	4	.176
	ed(1)	.007	1	.934
	ed(2)	.203	1	.652
	ed(3)	.835	1	.361
	ed(4)	5.773	1	.016
	retire	.013	1	.908
	gender	.214	1	.644
	tollfree	3.243	1	.072
	wireless	.668	1	.414
	tollmon	.000	1	.987
	cardmon	3.163	1	.075
	wiremon	1.084	1	.298
	pager	1.808	1	.179
	callwait	.266	1	.606
	forward	2.201	1	.138
	confer	2.568	1	.109
	custcat	.864	3	.834
	custcat(1)	.466	1	.495
	custcat(2)	.450	1	.502
	custcat(3)	.019	1	.889

Figura 349. Variáveis não no modelo (etapa 12 apenas)

Todas as variáveis deixadas de fora do modelo possuem estatísticas de pontuação com valores de significância maiores que 0.05 No entanto, os valores de relevância para *tollfree* e *cardmon*, embora não sejam menores que 0.05, são bastante próximos. Eles podem ser interessantes para buscar em mais estudos.

Médias de covariável

	Mean
age	41.684
marital	.505
address	11.551
income	77.535
ed(1)	.204
ed(2)	.287
ed(3)	.209
ed(4)	.234
employ	10.987
retire	.953
gender	.483
reside	2.331
tollfree	.526
equip	.614
callcard	.322
wireless	.704
longmon	11.723
tollmon	13.274
equipmon	14.220
cardmon	13.781
wiremon	11.584
multline	.525
voice	.696
pager	.739
internet	.632
callid	.519
callwait	.515
forward	.507
confer	.498
ebill	.629
custcat(1)	.266
custcat(2)	.217
custcat(3)	.281

Figura 350. Médias de covariável

Esta tabela exibe o valor médio de cada variável de preditor. Esta tabela é uma referência útil quando se olha para as parcelas de sobrevivência, que são construídas para os valores média. Note, no entanto, que o cliente "médio" não existe realmente quando você olha para os meios de variáveis indicadoras para preditores categóricos. Mesmo com todos os preditores de escala, é pouco provável que você encontre um cliente cujos valores covariados estejam todos próximos da média. Se você deseja ver a curva de sobrevivência para um determinado caso, você pode alterar os valores covariados em que a curva de sobrevivência é plotada na caixa de diálogo de Plots. Se você deseja ver a curva de sobrevivência para um determinado caso, você pode alterar os valores covariados em que a curva de sobrevivência é plotada no grupo de Plots do diálogo de Saída Avançada.

Curva de sobrevida

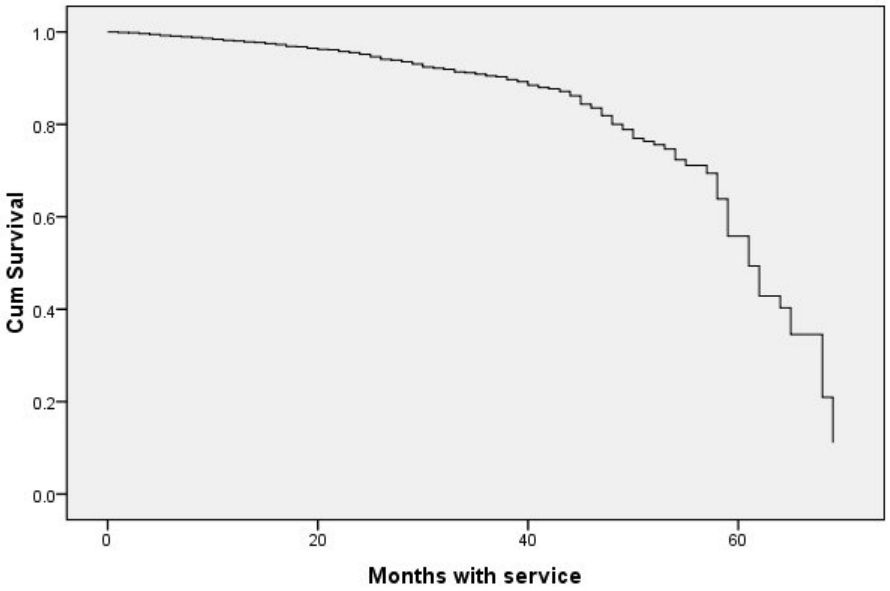


Figura 351. Curva de sobrevivência para cliente "médio"

A curva de sobrevivência básica é uma exibição visual do tempo previsto pelo modelo para o churn para o cliente "médio". O eixo horizontal mostra o tempo para o evento. O eixo vertical mostra a probabilidade de sobrevivência. Assim, qualquer ponto sobre a curva de sobrevivência mostra a probabilidade de que o cliente "médio" permaneça um cliente passado esse tempo. Passados 55 meses, a curva de sobrevivência torna-se menos suave. Há menos clientes que estiveram com a empresa por tanto tempo, por isso há menos informação disponível e, assim, a curva é bloqueada.

Curva De Risco

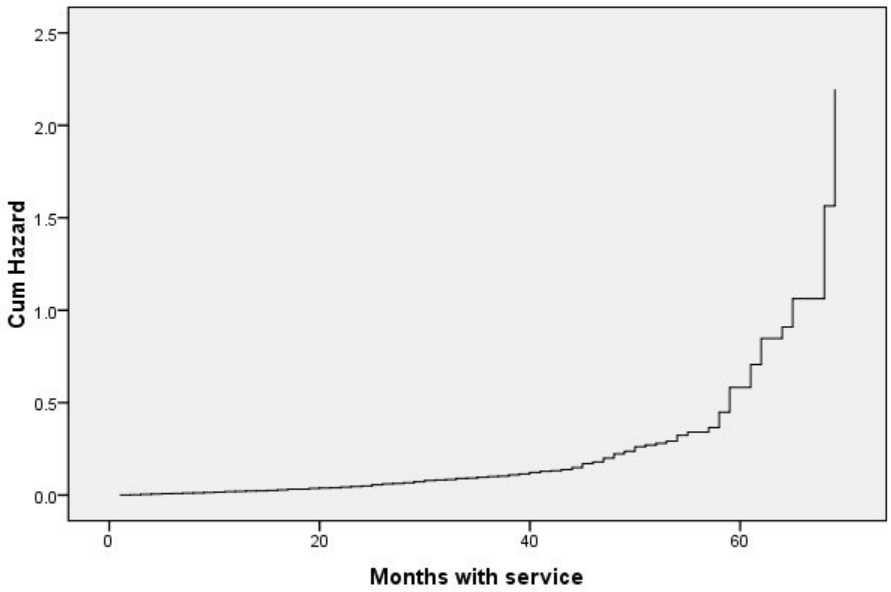


Figura 352. Curva de risco para cliente "médio"

A curva de risco básico é uma exibição visual do modelo acumulado-previsto para churn para o cliente "médio". O eixo horizontal mostra o tempo para o evento. O eixo vertical mostra o risco cumulativo, igual ao tronco negativo da probabilidade de sobrevivência. Passados 55 meses, a curva de risco, como a curva de sobrevivência, torna-se menos suave, pela mesma razão.

Avaliação

Os métodos de seleção stepwise garantem que o seu modelo terá apenas preditores "estatisticamente significativos", mas não garante que o modelo seja realmente bom em prever o alvo. Para isso, é necessário analisar registros marcados.

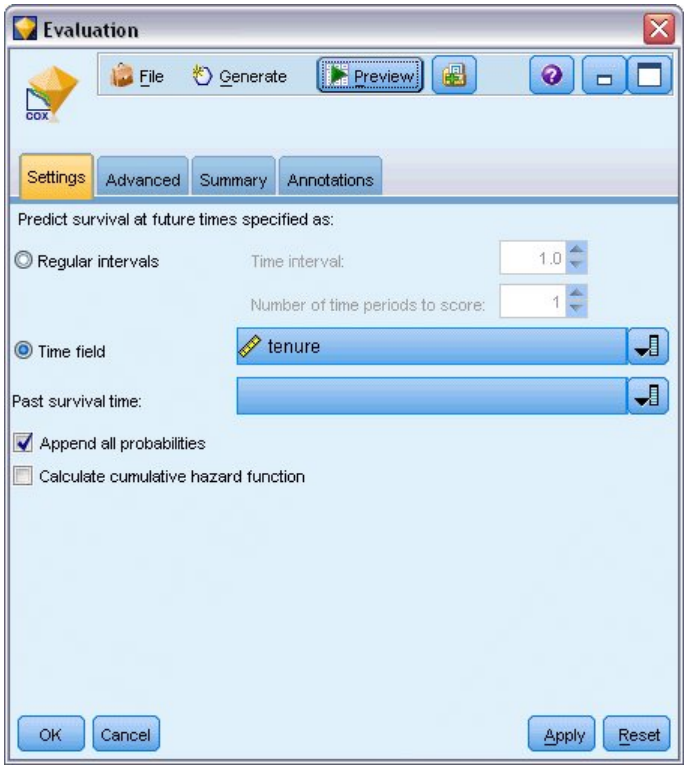


Figura 353. Cox pepita: Guia Configurações

1. Coloque o nugget modelo na tela e anexe-o ao nó de origem, abra o nugget e clique na aba Configurações.
2. Selecione **Campo de tempo** e especifique *tenure*. Cada registro será pontuado no seu comprimento de mandato.
3. Selecione **Append todas as probabilidades**.

   Isso cria pontuações usando 0.5 como o corte para saber se uma perda de clientes; se sua propensão para perda de clientes for maior que 0.5, elas serão pontuadas como uma perda de clientes. Não há nada mágico sobre esse número, e um corte diferente pode render resultados mais desejáveis. Para uma maneira de pensar em escolher um corte, use um nó de Avaliação.

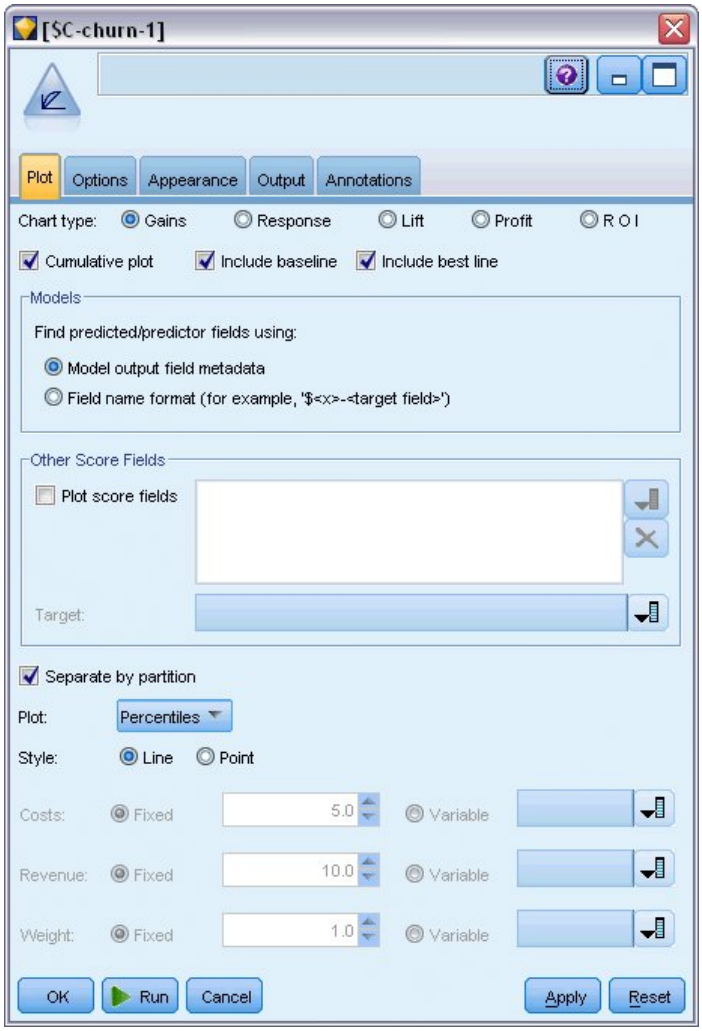


Figura 354. Nó de avaliação: Guia Plot

4. Conecte um nó de Avaliação ao nugget modelo; na guia Plot, selecione **Include melhor linha**.
5. Clique na guia **Opções**.

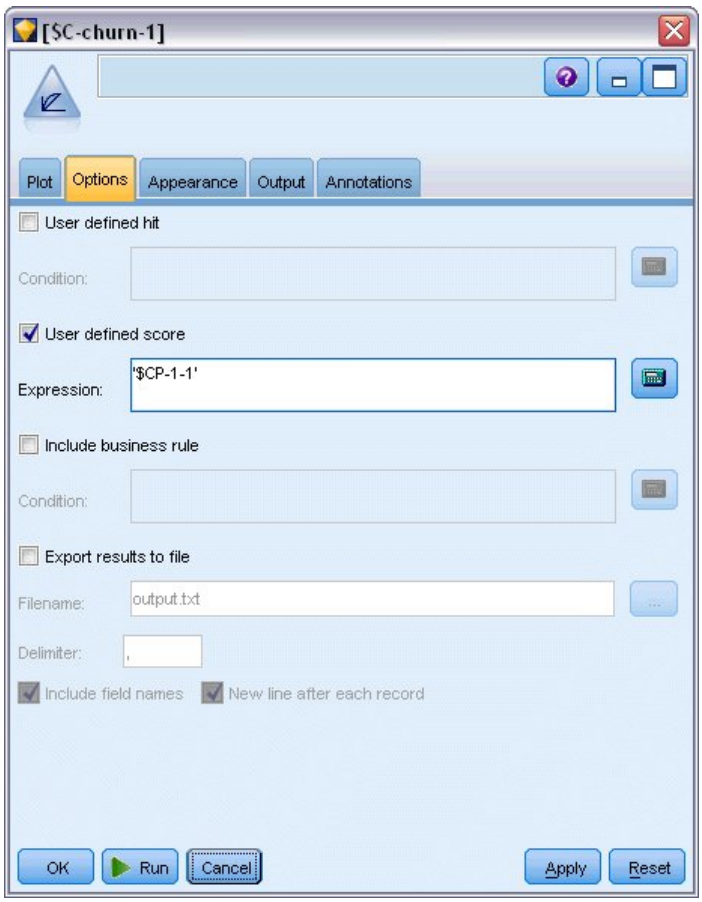


Figura 355. Nó de avaliação: guia Opções

6. Selecione **Pontuação definida pelo usuário** e digite `'$CP-1-1'` como a expressão Este é um campo gerado pelo modelo que corresponde à propensão ao churn.
7. Clique em **Executar** .

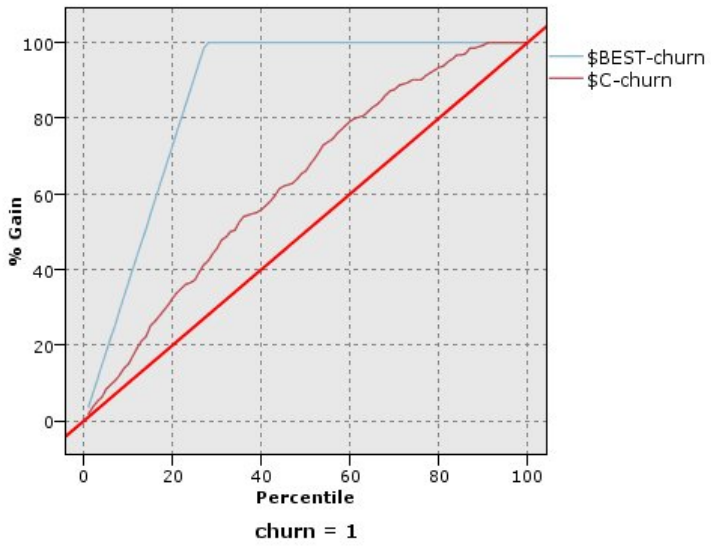


Figura 356. Gráfico de ganhos

O gráfico de ganhos acumulados mostra a porcentagem do número geral de casos em uma determinada categoria "ganhado" mirando uma porcentagem do número total de casos. Por exemplo, um ponto na curva está em (10%, 15%), significando que se você pontuar um dataset com o modelo e classificar todos os casos por propensão prevista para churn, você esperaria que o

top 10% contenha aproximadamente 15% de todos os casos que realmente levam a categoria *1* (churners). Da mesma forma, os 60% principais contêm aproximadamente 79.2% dos churners Se você selecionar 100% do dataset pontuado, você obtém todos os churners no dataset.

A linha diagonal é a curva "baseline"; se você selecionar 20% dos registros a partir do dataset pontuado ao acaso, você esperaria "ganhar" aproximadamente 20% de todos os registros que realmente levam a categoria *1*. Quanto mais longe acima da linha de base uma curva fica, maior o ganho. A "melhor" linha mostra a curva para um modelo "perfeito" que atribui uma pontuação de propensão de churn maior a cada churner do que cada não-churner. Você pode usar o gráfico de ganhos acumulativos para ajudar a escolher um corte de classificação escolhendo um percentual que corresponda a um ganho desejável e, em seguida, mapeando essa porcentagem para o valor de corte adequado.

O que constitui um ganho "desejável" depende do custo dos erros Tipo I e Tipo II. Ou seja, qual é o custo de classificar um churner como um não churner (Tipo I)? Qual é o custo de classificar um não churner como um churner (Tipo II)? Se a retenção do cliente for a principal preocupação, então você deseja diminuir seu erro Tipo I; no gráfico de ganhos acumulativos, isso pode corresponder ao aumento do atendimento ao cliente para os clientes nos 60% principais da propensão prevista de *1*, que captura 79.2% dos possíveis churners, mas custa tempo e recursos que poderiam ser gastos adquirindo novos clientes. Se baixar o custo de manutenção da sua base de clientes atual é a prioridade, então você quer diminuir o seu erro de Tipo II. No gráfico, isso pode corresponder a um aumento de atendimento ao cliente para os 20% principais, que captura 32.5% dos churners Usualmente, ambas são preocupações importantes, por isso é preciso escolher uma regra de decisão para classificar os clientes que dê o melhor mix de sensibilidade e especificidade.

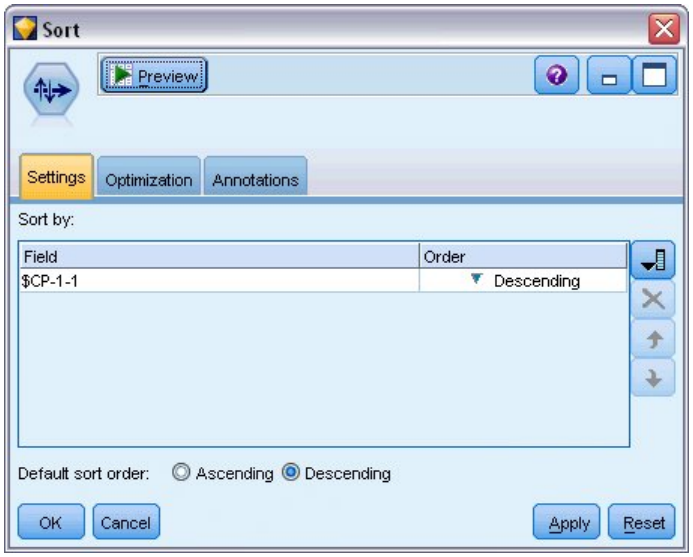
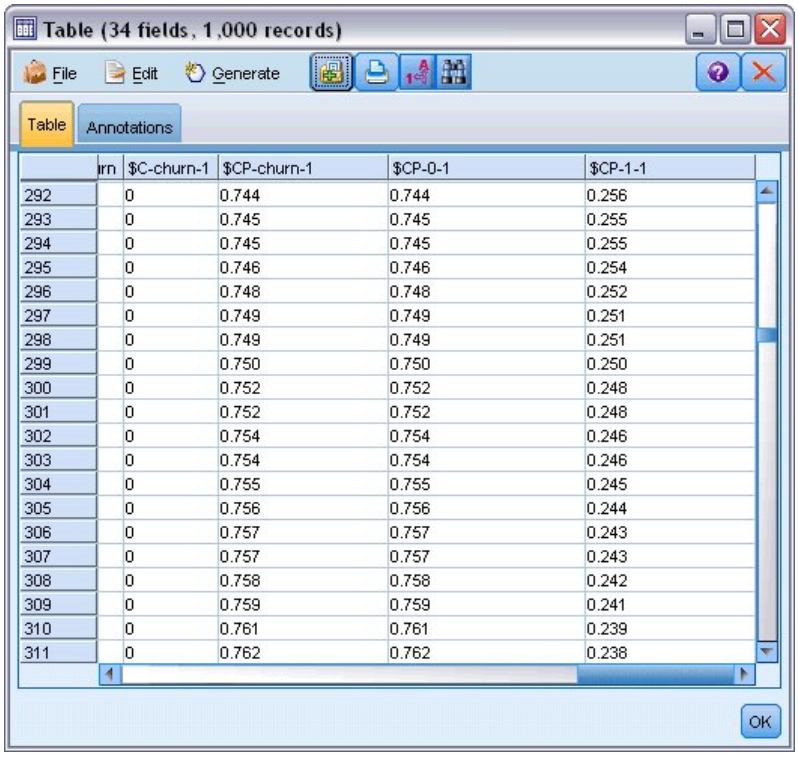


Figura 357. Nó de classificação: guia Configurações

8. Digamos que você decidiu que 45.6% é um ganho desejável, que corresponde a obter os 30% principais de registros. Para encontrar um cutoff de classificação apropriado, anexe um nó de Sort ao nugget modelo.
9. Na aba Configurações, escolha para classificar por *$CP-1-1* em ordem decrescente e clique em **OK**.



	ırn	$C-churn-1	$CP-churn-1	$CP-0-1	$CP-1-1
292		0	0.744	0.744	0.256
293		0	0.745	0.745	0.255
294		0	0.745	0.745	0.255
295		0	0.746	0.746	0.254
296		0	0.748	0.748	0.252
297		0	0.749	0.749	0.251
298		0	0.749	0.749	0.251
299		0	0.750	0.750	0.250
300		0	0.752	0.752	0.248
301		0	0.752	0.752	0.248
302		0	0.754	0.754	0.246
303		0	0.754	0.754	0.246
304		0	0.755	0.755	0.245
305		0	0.756	0.756	0.244
306		0	0.757	0.757	0.243
307		0	0.757	0.757	0.243
308		0	0.758	0.758	0.242
309		0	0.759	0.759	0.241
310		0	0.761	0.761	0.239
311		0	0.762	0.762	0.238

Figura 358. Tabela

10. Conecte um nó da Tabela ao nó Sort.
11. Abra o nó da Tabela e clique em **Executar**.

Rolando a saída para baixo, você vê que o valor de *$CP-1-1* é 0.248 para o 300º registro. O uso do site 0.248 como ponto de corte de classificação deve resultar em aproximadamente 30% dos clientes classificados como churners, capturando aproximadamente 45% do total real de churners.

Rastreamento do Número Esperado de Clientes Retidos

Uma vez satisfeito com um modelo, deseja-se acompanhar o número esperado de clientes no dataset que são retidos nos próximos dois anos. Os valores nulos, que são clientes cujo mandato total (tempo futuro + *tenure*) cai além da faixa de tempo de sobrevivência nos dados utilizados para treinar o modelo, apresentam um desafio interessante. Uma forma de lidar com eles é criar dois conjuntos de predições, uma em que os valores nulos são assumidos de terem churgado, e outro em que se assumem que foram retidos. Desta forma é possível estabelecer limites superiores e inferiores sobre o número esperado de clientes retidos.

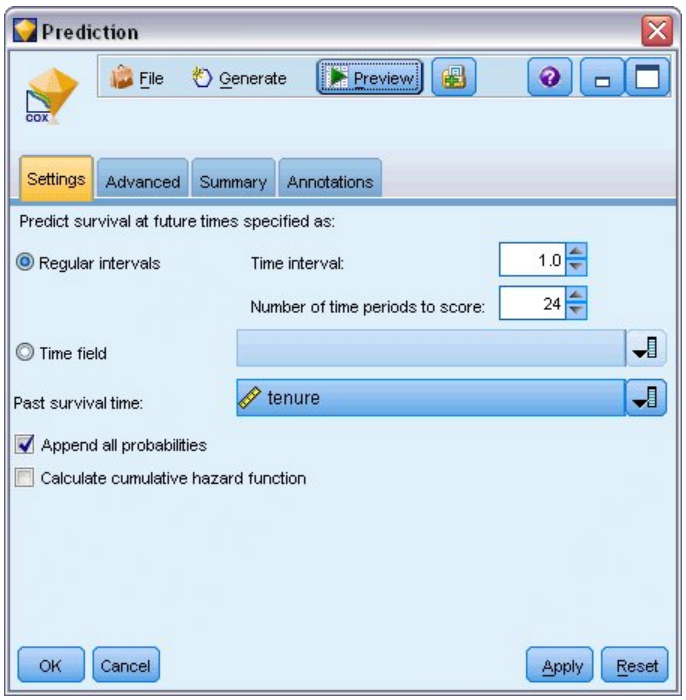


Figura 359. Cox pepita: Guia Configurações

1. Clique duas vezes no nugget de modelo na paleta de Modelos (ou copie e cole o nugget na tela do fluxo) e anexe o novo nugget no nó Fonte.
2. Abra o nugget para a aba Configurações.
3. Certifique-se de que **Intervalos Regulares** esteja selecionado e especifique `1.0` como o intervalo de tempo e 24 como o número de períodos para pontuação. Isso especifica que cada registro será pontuado para cada um dos 24 meses seguintes.
4. Selecione *tenure* como o campo para especificar o tempo de sobrevivência passado. O algoritmo de pontuação levará em conta o comprimento do tempo de cada cliente como cliente da empresa.
5. Selecione **Append todas as probabilidades**.

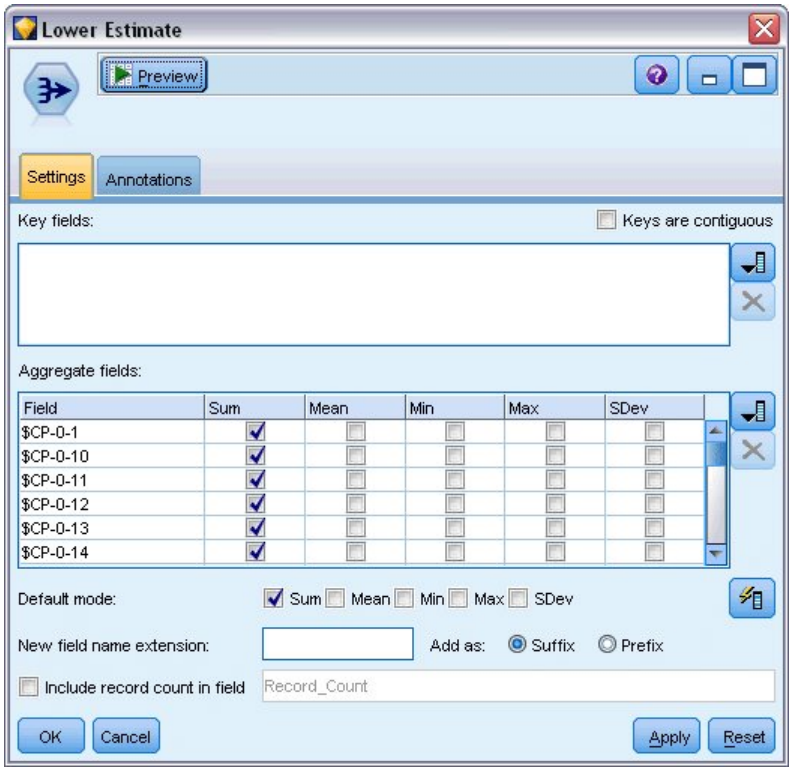


Figura 360. Nó agregado: guia Configurações

6. Conecte um nó Agregado ao nugget do modelo; na aba Configurações, desmarque **Mean** como um modo padrão.
7. Selecione *$CP-0-1* através do *$CP-0-24*, os campos do formulário *$CP-0-n*, como os campos a agregar. Isso é mais fácil se, no diálogo Selecionar Campos, você classificar os campos por Nome (isto é, ordem alfabética).
8. Desmarque **Include contagem de registros em campo**.
9. Clique em **OK**. Este nó cria as predições "inferiores ligadas".

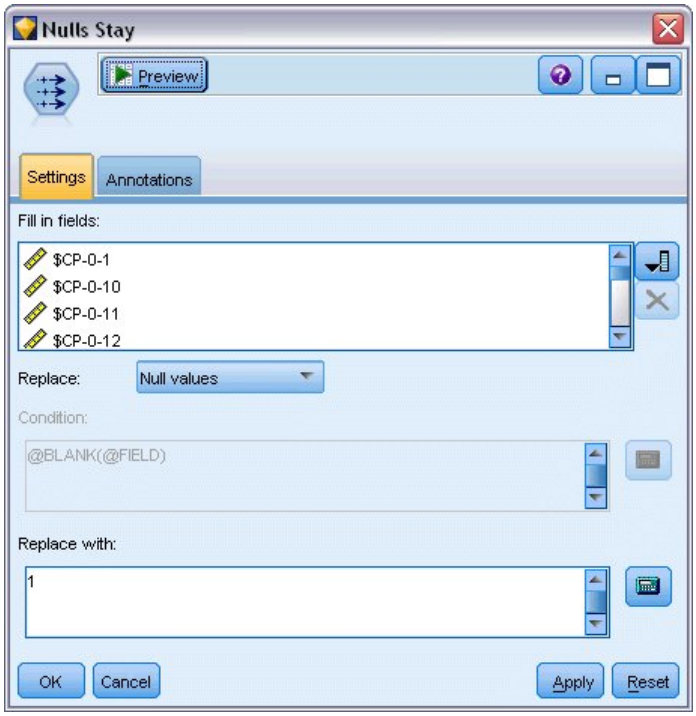


Figura 361. Nó do preenchimento: guia Configurações

10. Conecte um nó Filler ao nugget Coxreg ao qual acabamos de anexar o nó do Agregado; na guia Configurações, selecione *$CP-0-1* através do *$CP-0-24*, os campos do formulário *$CP-0-n*, como os campos a preencher. Isso é mais fácil se, no diálogo Selecionar Campos, você classificar os campos por Nome (isto é, ordem alfabética).
11. Escolha para substituir **Valores Null** com o valor 1.
12. Clique em **OK**.

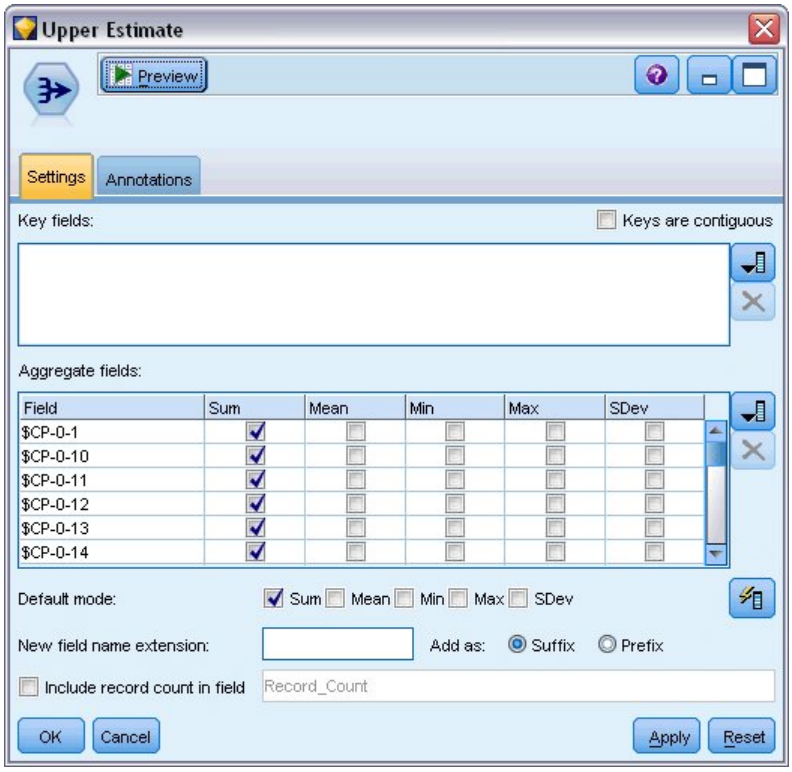


Figura 362. Nó agregado: guia Configurações

13. Conecte um nó Agregado ao nó Filler; na aba Configurações, desmarque **Mean** como um modo padrão.
14. Selecione *$CP-0-1* através do *$CP-0-24*, os campos do formulário *$CP-0-n*, como os campos a agregar. Isso é mais fácil se, no diálogo Selecionar Campos, você classificar os campos por Nome (isto é, ordem alfabética).
15. Desmarque **Include contagem de registros em campo**.
16. Clique em **OK**. Este nó cria as predições de "limite superior".

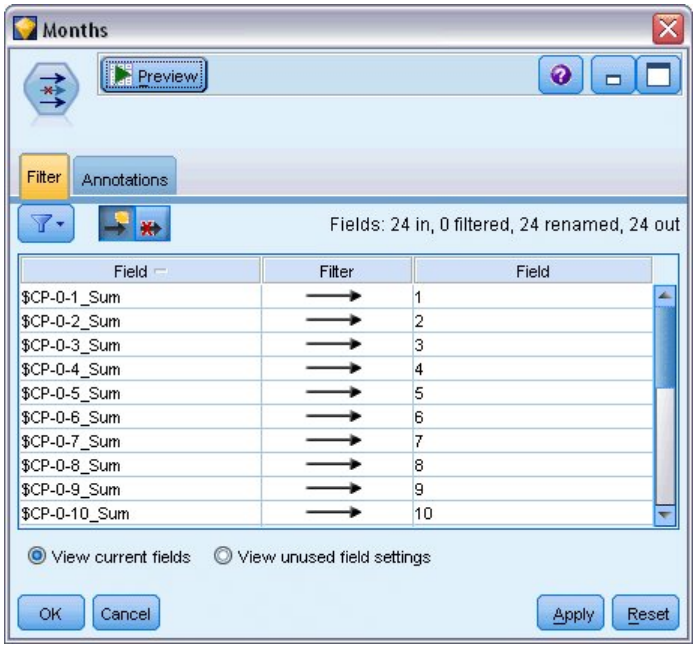


Figura 363. Nó do filtro: guia Configurações

17. Fixe um nó Append para os dois nós Agregados, em seguida, anexe um nó Filtro ao nó Append.
18. Na guia Configurações do nó Filtro, renomear os campos para *1* através de *24*. Através do uso de um nó Transpose, esses nomes de campo se tornarão valores para o eixo *x*em gráficos a jusante.

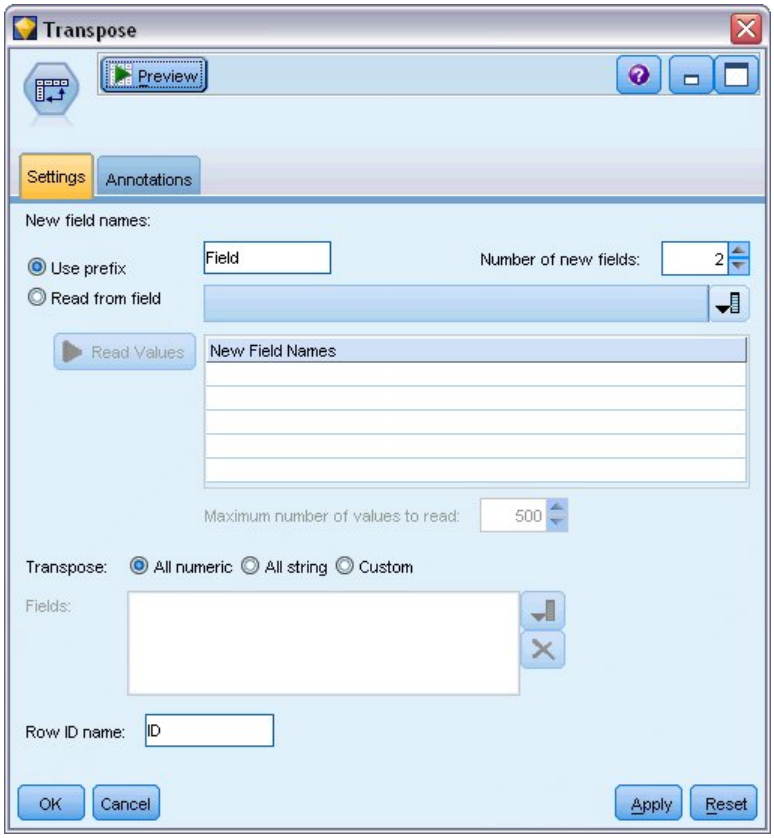


Figura 364. Transpor nó: guia Configurações

19. Conecte um nó Transpose ao nó Filtro.
20. Digite 2 como o número de novos campos

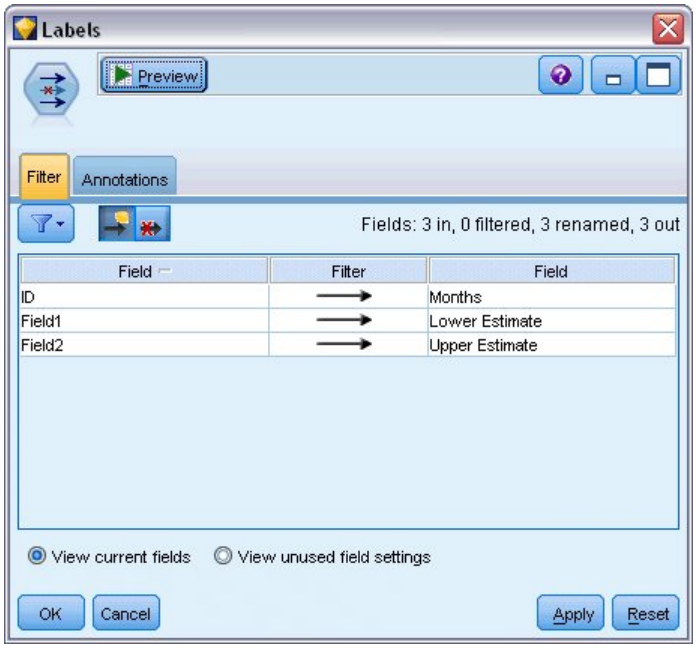


Figura 365. Nó do filtro: guia Filtro

21. Conecte um nó Filtro ao nó Transpose.
22. Na guia Configurações do nó Filtro, renomear *ID* para *Months*, *Field1* para *Lower Estimate*, e *Field2* para *Estiagem Superior*.

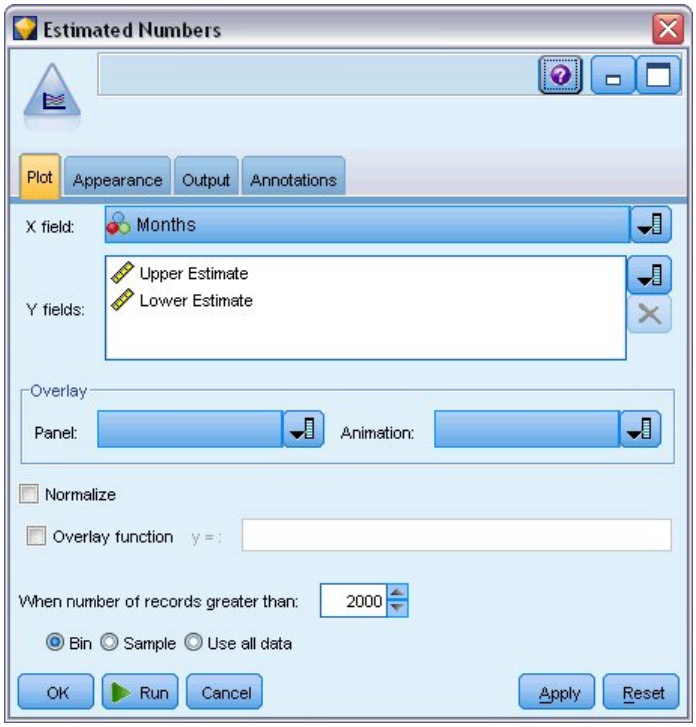


Figura 366. Nó multiplot: guia Plot

23. Conecte um nó Multiplot ao nó Filtro.
24. Na guia Plot, *Months* como campo X, *Lower Estimate* e *Upper Estimate* como os campos Y.

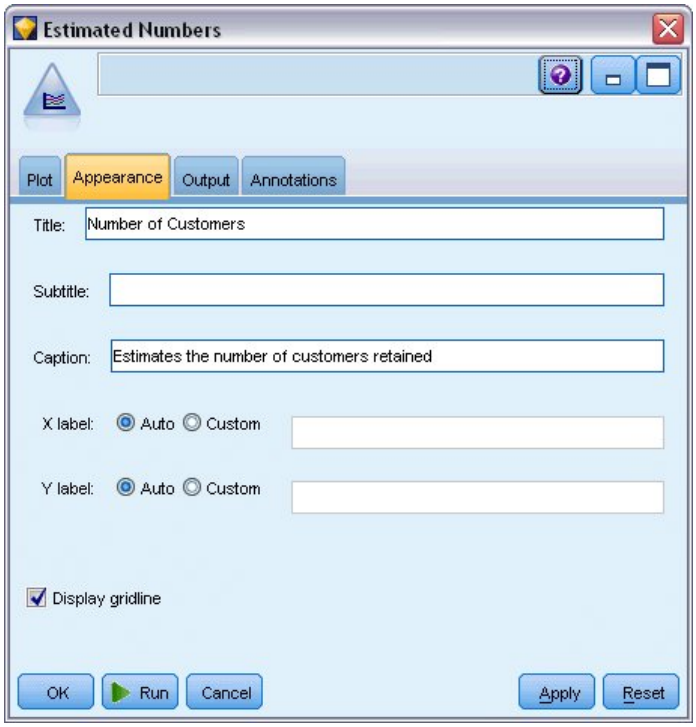


Figura 367. Nó multiplot: guia Aparência

25. Clique na guia Aparência.
26. Digite `Number of Customers` como o título..
27. Digite `Estimates the number of customers retained` como a legenda
28. Clique em **Executar** .

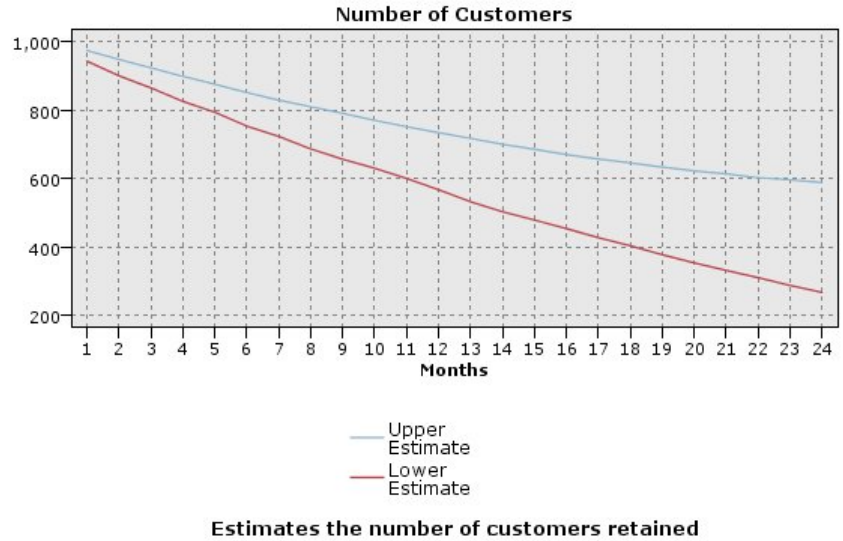


Figura 368. Multiplot estimando o número de clientes retidos

Os limites superiores e inferiores sobre o número estimado de clientes retidos são plotados. A diferença entre as duas linhas é o número de clientes pontuados como nulo e, portanto, cujo status é altamente incerto. Com o tempo, o número desses clientes aumenta. Após 12 meses, você pode esperar reter entre 601 e 735 dos clientes originais no dataset; depois de 24 meses, entre 288 e 597.

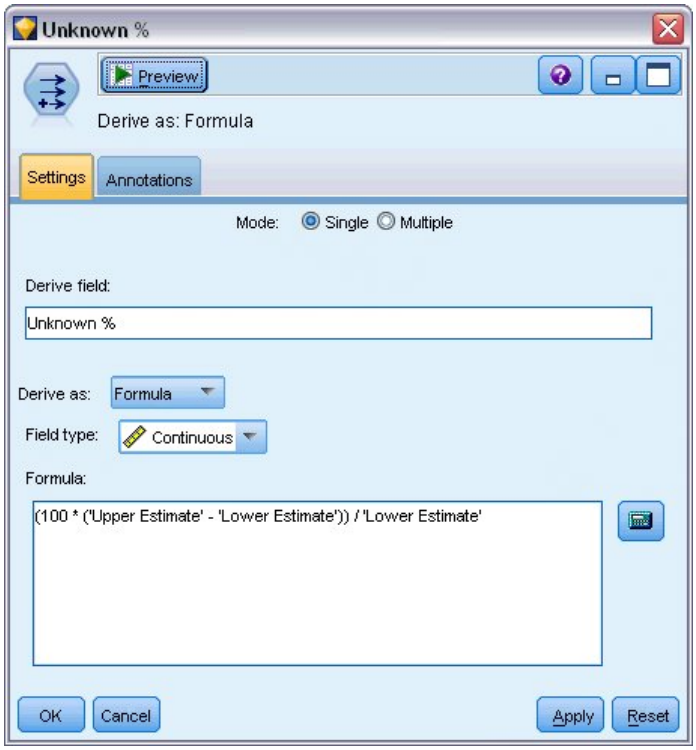


Figura 369. Derivar nó: guia Configurações

29. Para obter outro olhar sobre como são incertas as estimativas do número de clientes retidos são, anexar um nó de Derivação ao nó Filtro.
30. Na guia Configurações do nó do Derivar, digite *Unknown%* como o campo derivar.
31. Selecione **Continuous** como o tipo de campo.
32. Digite `(100 * ('Upper Estimate' - 'Lower Estimate')) / 'Lower Estimate'` como a fórmula *% desconhecida* é o número de clientes "em dúvida" como uma porcentagem da estimativa inferior.
33. Clique em **OK**.

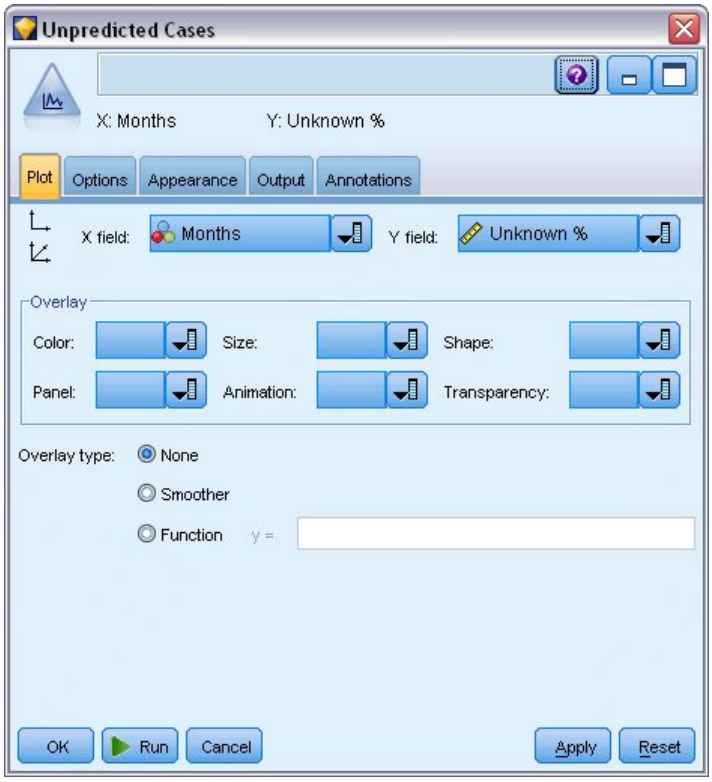


Figura 370. Nó da trama: guia Plot

34. Conecte um nó de Plot ao nó do Derivar.
35. Na guia Plot do nó do Plot, selecione *Months* como o campo X e *Unknown%* como o campo Y.
36. Clique na guia **Aparência** .

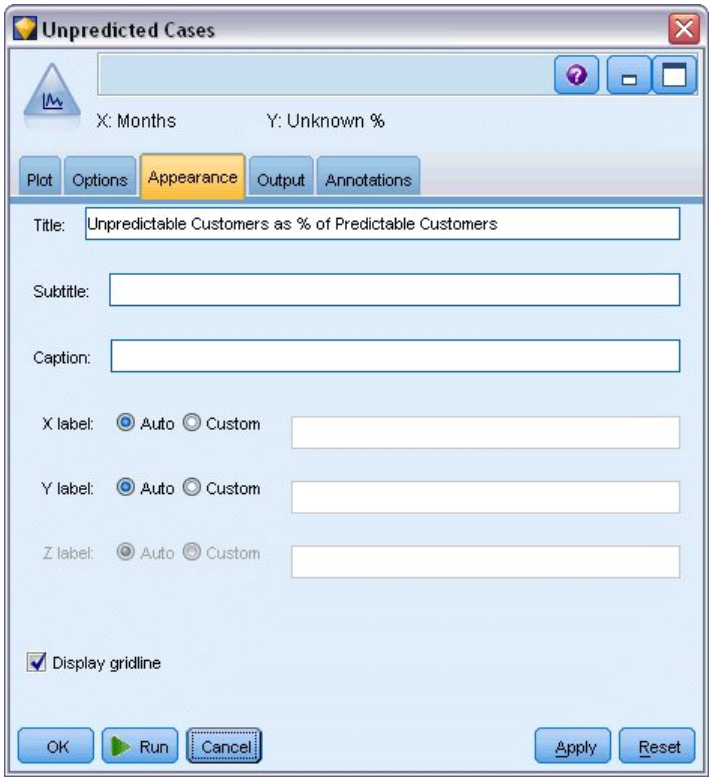


Figura 371. Nó da trama: guia Aparência

37. Digite `Unpredictable Customers as % of Predictable Customers` como o título..

38. Execute o nó.

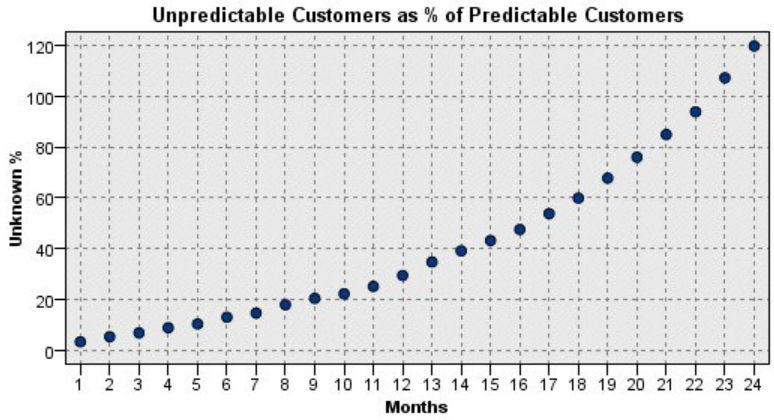


Figura 372. Trama de clientes imprevisíveis

Através do primeiro ano, a porcentagem de clientes imprevisíveis aumenta a uma taxa bastante linear, mas a taxa de aumento explode durante o segundo ano até que, até o mês 23, o número de clientes com valores nulos supere o número esperado de clientes retidos.

Escoragem

Uma vez satisfeito com um modelo, você quer pontuar os clientes para identificar os indivíduos mais propensos a se churn dentro do próximo ano, por trimestre.

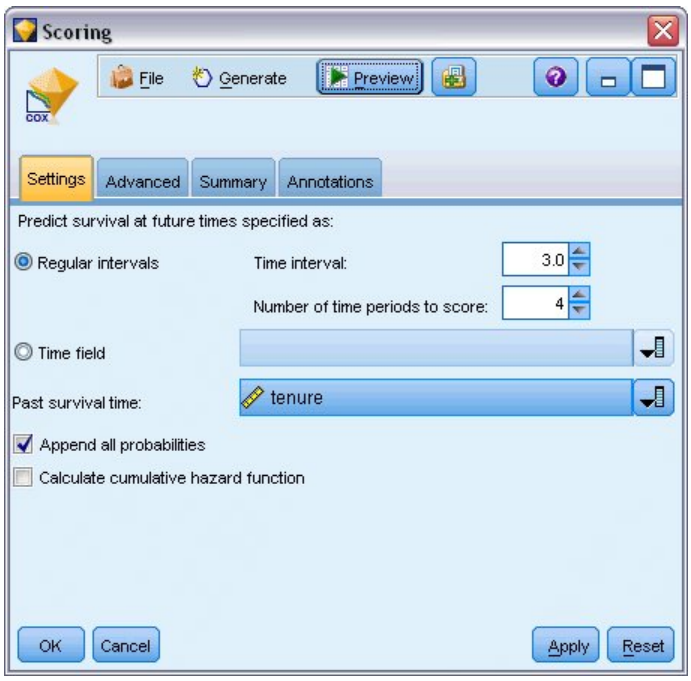


Figura 373. Coxreg nugget: guia Configurações

1. Conecte um nugget de terceiro modelo ao nó Fonte e abra o nugget modelo.
2. Certifique-se de que **Intervalos Regulares** esteja selecionado e especifique `3.0` como o intervalo de tempo e 4 como o número de períodos para pontuação. Isso especifica que cada registro será pontuado para os quatro trimestres seguintes.
3. Selecione *tenure* como o campo para especificar o tempo de sobrevivência passado. O algoritmo de pontuação levará em conta o comprimento do tempo de cada cliente como cliente da empresa.

4. Selecione **Append todas as probabilidades**. Esses campos extras tornarão mais fácil classificar os registros para visualização em uma tabela.

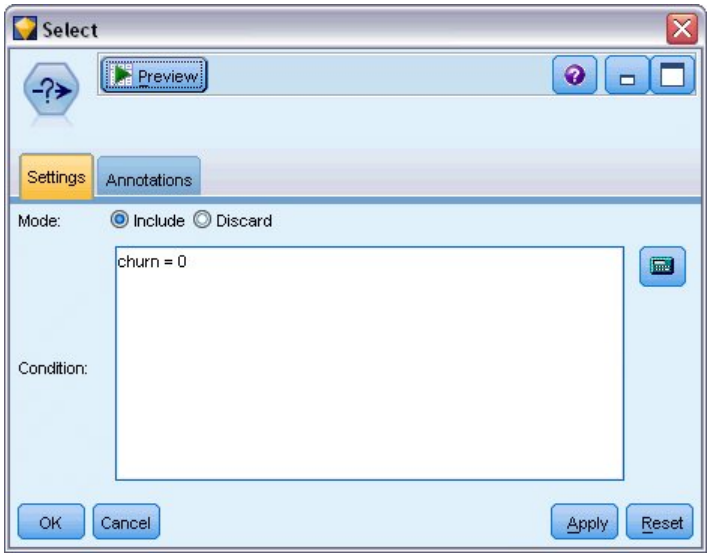


Figura 374. Selecionar nó: guia Configurações

5. Anexe um nó de Seleção ao nugget do modelo; na guia Configurações, digite `churn=0` como a condição Isso elimita os clientes que já se agilearam da tabela de resultados.

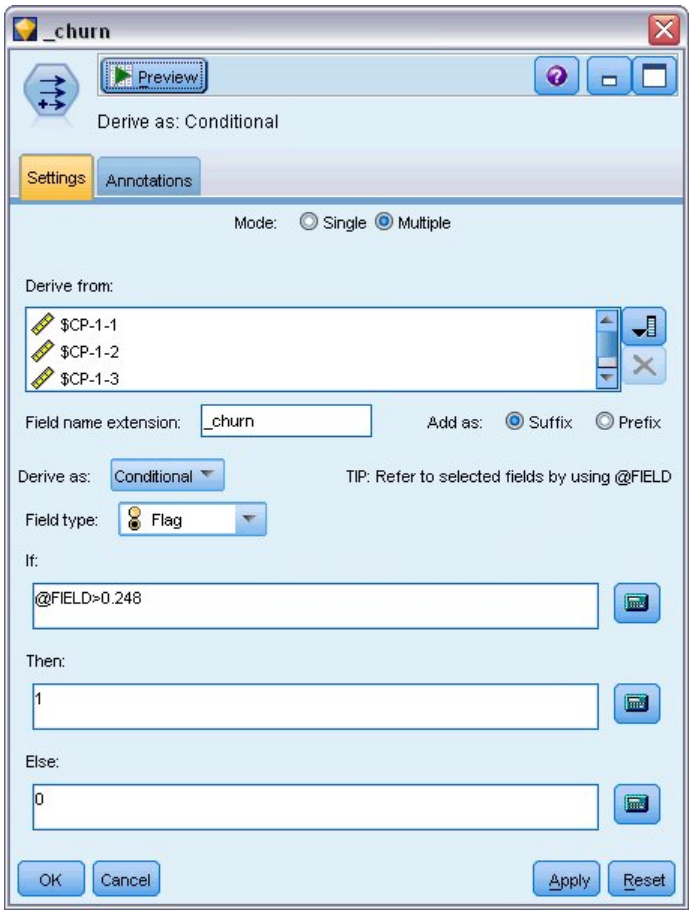


Figura 375. Derivar nó: guia Configurações

6. Anexar um nó do Derivar ao nó Select; na guia Configurações, selecione **Vários** como o modo.
7. Escolha derivar de *$CP-1-1* até *$CP-1-4*, os campos do formulário *$CP-1-n*e digite `_churn` como o sufixo a ser incluído. Isso é mais fácil se, no diálogo Selecionar Campos, você classificar os campos por Nome (isto é, ordem alfabética).

8. Opte por derivar o campo como um **Condicionado**.
9. Selecione **Bandeira** como o nível de medição.
10. Digite `@FIELD>0.248` como a condição **If** .. Lembre-se que este foi o corte de classificação identificado durante a Avaliação.
11. Digite `1` como a expressão **Then** .
12. Digite `0` como a expressão **Else** ..
13. Clique em **OK**.

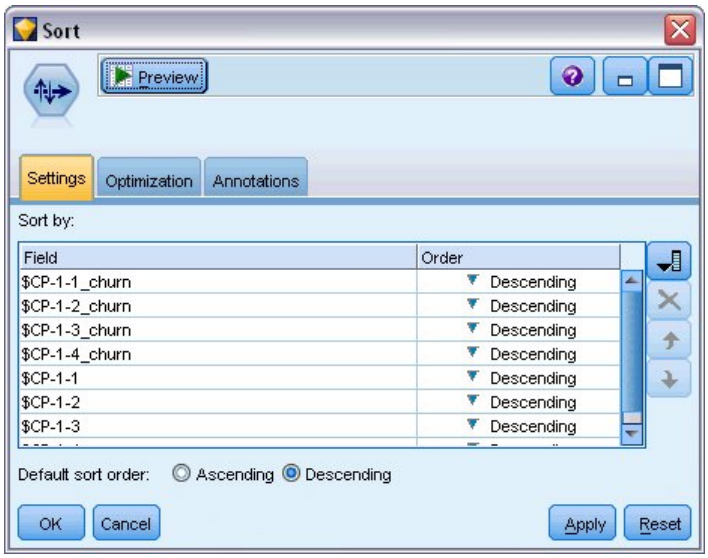


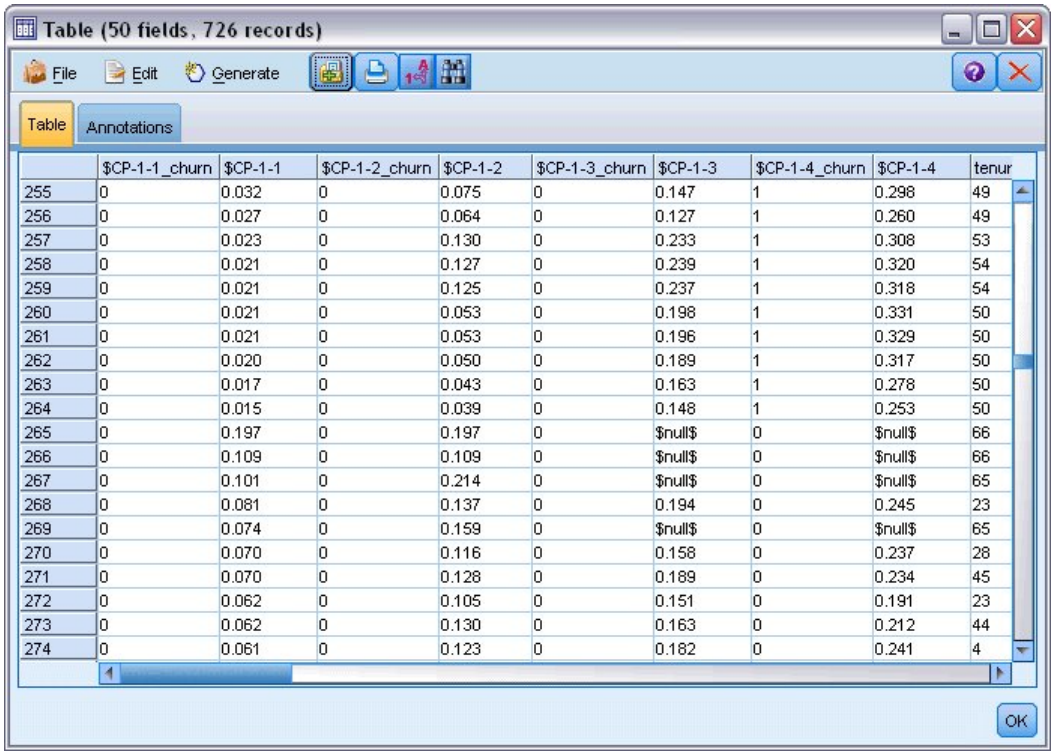
Figura 376. Nó de classificação: guia Configurações

14. Conecte um nó Sort ao nó do Derivação; na aba Configurações, escolha classificar por *$CP-1-1_churn* por meio de *$CP-1-4-churn* e depois *$CP-1-1* através de *$CP-1-4*, tudo em ordem decrescente. Os clientes que estão previstos para churn aparecerão no topo.

Figura 377. Nó Reordem de Campo: guia Reordenar

15. Conecte um nó Reordem de Campo ao nó Sort; na guia Reordenar, escolha colocar *$CP-1-1_churn* por meio do *$CP-1-4* na frente dos outros campos. Isso simplesmente torna a tabela de resultados

mais fácil de ler, e assim é opcional. Você precisará utilizar os botões para mover os campos para a posição mostrada na figura.

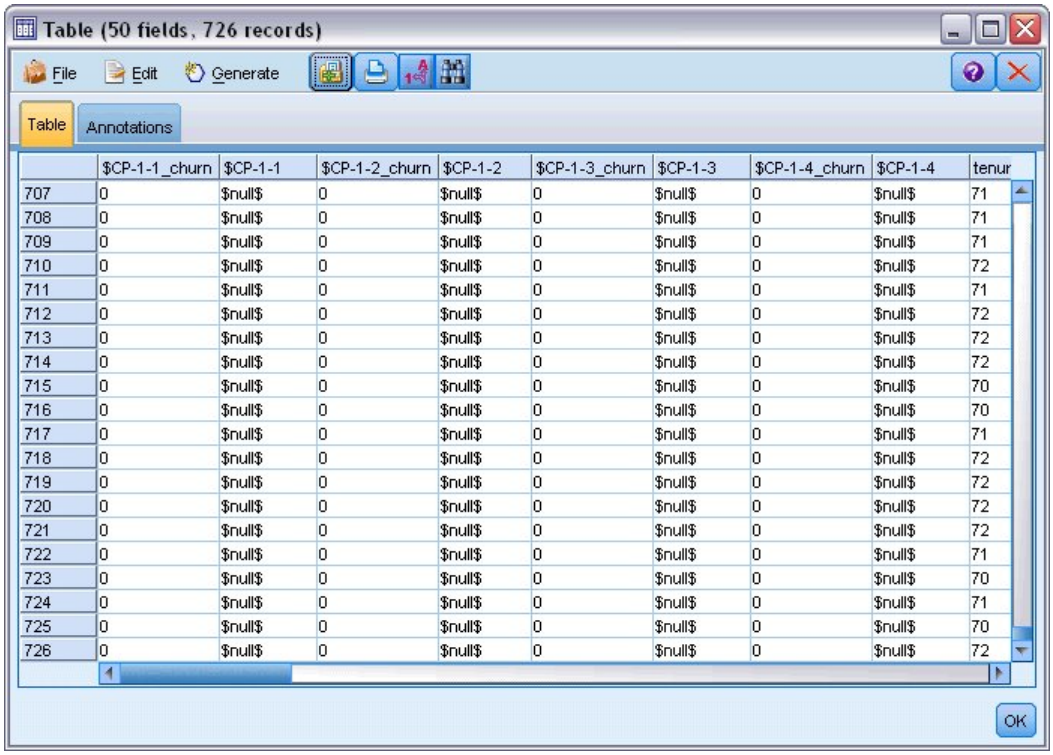


	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
255	0	0.032	0	0.075	0	0.147	1	0.298	49
256	0	0.027	0	0.064	0	0.127	1	0.260	49
257	0	0.023	0	0.130	0	0.233	1	0.308	53
258	0	0.021	0	0.127	0	0.239	1	0.320	54
259	0	0.021	0	0.125	0	0.237	1	0.318	54
260	0	0.021	0	0.053	0	0.198	1	0.331	50
261	0	0.021	0	0.053	0	0.196	1	0.329	50
262	0	0.020	0	0.050	0	0.189	1	0.317	50
263	0	0.017	0	0.043	0	0.163	1	0.278	50
264	0	0.015	0	0.039	0	0.148	1	0.253	50
265	0	0.197	0	0.197	0	$null$	0	$null$	66
266	0	0.109	0	0.109	0	$null$	0	$null$	66
267	0	0.101	0	0.214	0	$null$	0	$null$	65
268	0	0.081	0	0.137	0	0.194	0	0.245	23
269	0	0.074	0	0.159	0	$null$	0	$null$	65
270	0	0.070	0	0.116	0	0.158	0	0.237	28
271	0	0.070	0	0.128	0	0.189	0	0.234	45
272	0	0.062	0	0.105	0	0.151	0	0.191	23
273	0	0.062	0	0.130	0	0.163	0	0.212	44
274	0	0.061	0	0.123	0	0.182	0	0.241	4

Figura 378. Tabela mostrando pontuações do cliente

16. Conecte um nó da Tabela ao nó de Reordem de Campo e execute-o.

São esperados 264 clientes para churn até o final do ano, 184 até o final do terceiro trimestre, 103 pelo segundo, e 31 no primeiro. Nota que, dada a dois clientes, aquele com maior propensão ao churn no primeiro trimestre não tem necessariamente uma propensão maior ao churn em trimestres posteriores; por exemplo, ver os registros 256 e 260. Isso é provável devido à forma da função de risco para os meses seguintes ao atual mandato do cliente; por exemplo, os clientes que se uniram por causa de uma promoção podem ser mais propensos a trocar mais cedo do que os clientes que se juntaram por causa de uma recomendação pessoal, mas se não o fizer então podem realmente ser mais leais para o seu mandato remanescente. Você pode querer re-classificar os clientes para obter visões diferentes dos clientes mais propensos a churn.



	$CP-1-1_churn	$CP-1-1	$CP-1-2_churn	$CP-1-2	$CP-1-3_churn	$CP-1-3	$CP-1-4_churn	$CP-1-4	tenur
707	0	$null$	0	$null$	0	$null$	0	$null$	71
708	0	$null$	0	$null$	0	$null$	0	$null$	71
709	0	$null$	0	$null$	0	$null$	0	$null$	71
710	0	$null$	0	$null$	0	$null$	0	$null$	72
711	0	$null$	0	$null$	0	$null$	0	$null$	71
712	0	$null$	0	$null$	0	$null$	0	$null$	72
713	0	$null$	0	$null$	0	$null$	0	$null$	72
714	0	$null$	0	$null$	0	$null$	0	$null$	72
715	0	$null$	0	$null$	0	$null$	0	$null$	70
716	0	$null$	0	$null$	0	$null$	0	$null$	70
717	0	$null$	0	$null$	0	$null$	0	$null$	71
718	0	$null$	0	$null$	0	$null$	0	$null$	72
719	0	$null$	0	$null$	0	$null$	0	$null$	72
720	0	$null$	0	$null$	0	$null$	0	$null$	72
721	0	$null$	0	$null$	0	$null$	0	$null$	72
722	0	$null$	0	$null$	0	$null$	0	$null$	71
723	0	$null$	0	$null$	0	$null$	0	$null$	70
724	0	$null$	0	$null$	0	$null$	0	$null$	71
725	0	$null$	0	$null$	0	$null$	0	$null$	70
726	0	$null$	0	$null$	0	$null$	0	$null$	72

Figura 379. Tabela mostrando clientes com valores nulos

Na parte inferior da tabela estão os clientes com valores nulos previstos. São clientes cujo total de arrendamento (tempo futuro + *tenure*) cai além da faixa de tempos de sobrevivência nos dados utilizados para treinar o modelo.

Resumo

Usando a regressão Cox, você encontrou um modelo aceitável para o tempo de rotatividade, traçou o número esperado de clientes retidos nos próximos dois anos e identificou os clientes individuais com maior probabilidade de rotatividade no próximo ano. Observe que, embora este seja um modelo aceitável, pode não ser o melhor modelo. O ideal você deve, pelo menos, comparar este modelo, obtido usando o método Forward stepwise, com um criado usando o método stepwise de Backward.

Explicações sobre as bases matemáticas dos métodos de modelagem utilizados em IBM SPSS Modelador estão listadas no *IBM SPSS Modelador Algorithms Guide*.

Capítulo 27. Análise Da Cesta De Mercado (Regra Induction/C5.0)

Este exemplo trata de dados fictícios descrevendo o conteúdo de cestas de supermercados (ou seja, coleções de itens comprados em conjunto) mais os dados pessoais associados do comprador, que podem ser adquiridos através de um esquema de cartão de fidelidade. O objetivo é descobrir grupos de clientes que compram produtos similares e que possam ser caracterizados demograficamente, como por exemplo, por idade, renda, entre outros.

Este exemplo ilustra duas fases de mineração de dados:

- Modelagem de regra de associação e um display web revelando links entre itens comprados
- C5.0 indução de indução de perfis os compradores de grupos de produtos identificados

Nota: Este aplicativo não faz uso direto da modelagem preditiva, portanto, não há medição de precisão para os modelos resultantes e nenhuma distinção de treinamento / teste associado no processo de mineração de dados.

Este exemplo usa o fluxo denominado *baskrule*, que faz referência ao arquivo de dados denominado *BASKETS1n*. Esses arquivos estão disponíveis a partir do diretório *Demos* de qualquer instalação IBM SPSS Modelador . Isso pode ser acessado a partir do grupo do programa IBM SPSS Modelador no menu Iniciar do Windows. O arquivo *baskrule* está no diretório *streams* .

Acessando os Dados

Usando um nó do File Variable, conecte-se ao dataset *BASKETS1n*, selecionando para ler nomes de campo do arquivo. Conecte um nó Tipo à fonte de dados e, em seguida, conecte o nó a um nó da Tabela. Configure o nível de medição do campo *cardid* para *Typeless* (pois cada ID do cartão de fidelidade ocorre apenas uma vez no dataset e, portanto, não pode ser de nenhum uso na modelagem). Selecione *Nominal* como o nível de medição para o campo *sex* (isto é para garantir que o algoritmo de modelagem Apriori não tratará *sexo* como uma bandeira).

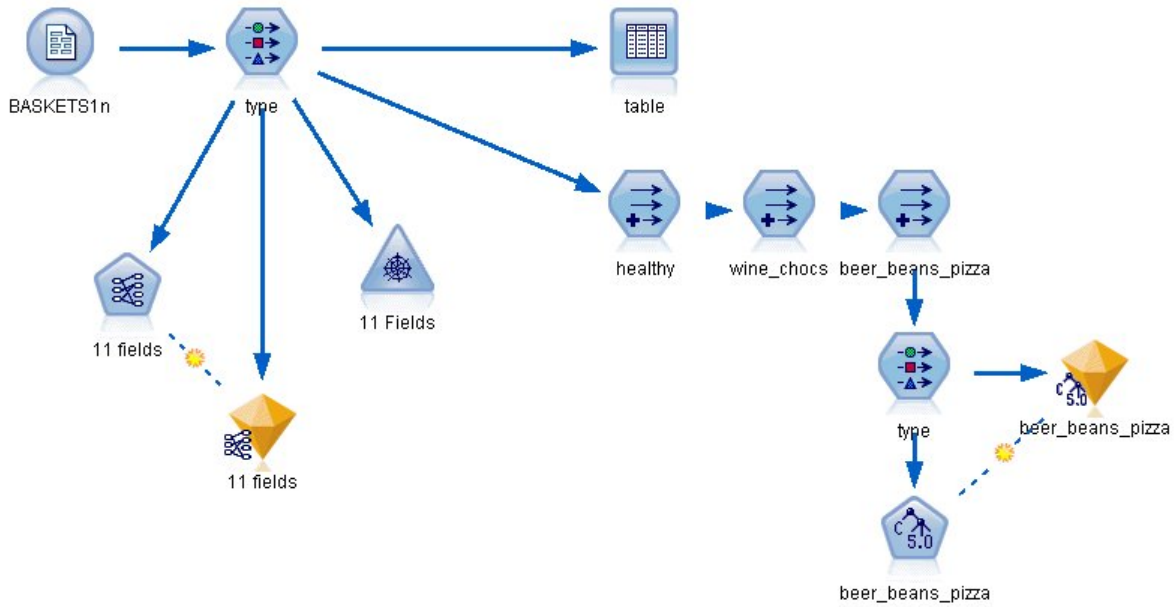


Figura 380. fluxo de baskrule

Agora execute o fluxo para instanciar o nó Type e exibir a tabela. O dataset contém 18 campos, com cada registro representando uma cesta.

Os 18 campos são apresentados nas seguintes rubricas.

Resumo da Cesta:

- *cardid*. Identificador de cartão de fidelidade para compra de cliente esta cesta.
- *valor*. Preço de compra total da cesta.
- *método*. Método de pagamento para cesta.

Detalhes pessoais do titular do cartão:

- *Sexo*
- *proprietário*. Se o titular do cartão é ou não proprietário de casa.
- *INCOME*
- *Idade*

Conteúdos da cesta-bandeiras para presença de categorias de produtos:

- *infrutíveg*
- *carne fresca*
- *laticínio*
- *vegetais enlatados*
- *carne enlatada*
- *frozenrefeição*
- *cerveja*
- *vinho*
- *refrigerante*
- *peixe*
- *confeitaria*

Descobrindo afinidades em conteúdos da cesta

Primeiro, você precisa adquirir um quadro geral de afinidades (associações) no conteúdo da cesta utilizando Apriori para produzir regras de associação. Selecione os campos a serem usados neste processo de modelagem, editando o nó Type e configurando a função de todas as categorias de produto para *Both* e todas as outras funções para *Nenhum*. (*Ambos* significa que o campo pode ser uma entrada ou uma saída do modelo resultante.)

Nota: Você pode configurar opções para vários campos usando Shift-click para selecionar os campos antes de especificar uma opção a partir das colunas.

Figura 381. Seleção de campos para modelagem

Uma vez que você tenha campos especificados para modelagem, anexe um nó Apriori ao nó Type, edite-o, selecione a opção **Somente valores verdadeiros para sinalizadores**e clique em executar no nó Apriori. O resultado, um modelo na guia Models na parte superior direita da janela de gerentes, contém regras de associação que você pode visualizar usando o menu de contexto e selecionando **Procurar**.

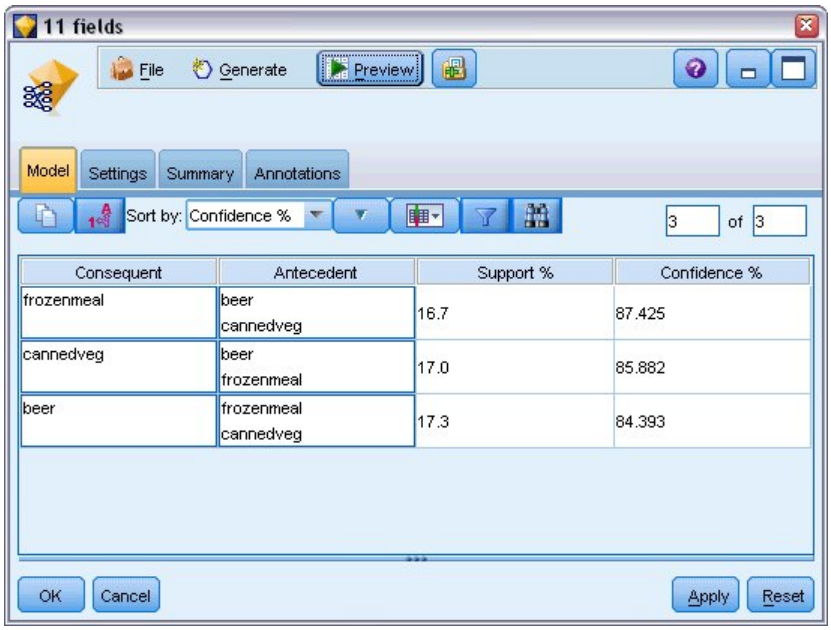


Figura 382. Regras de associação

Essas regras mostram uma variedade de associações entre refeições congeladas, legumes enlatados e cerveja. A presença de regras de associação de duas vias, tais como:

```
frozenmeal -> beer
beer -> frozenmeal
```

sugere que um display web (que mostra apenas associações de duas vias) possa destacar alguns dos padrões nestes dados.

Conecte um nó da Web ao nó Type, edite o nó da Web, selecione todos os campos de conteúdos da cesta, selecione **Mostrar somente as verdadeiras bandeiras**e clique em executar no nó da Web.

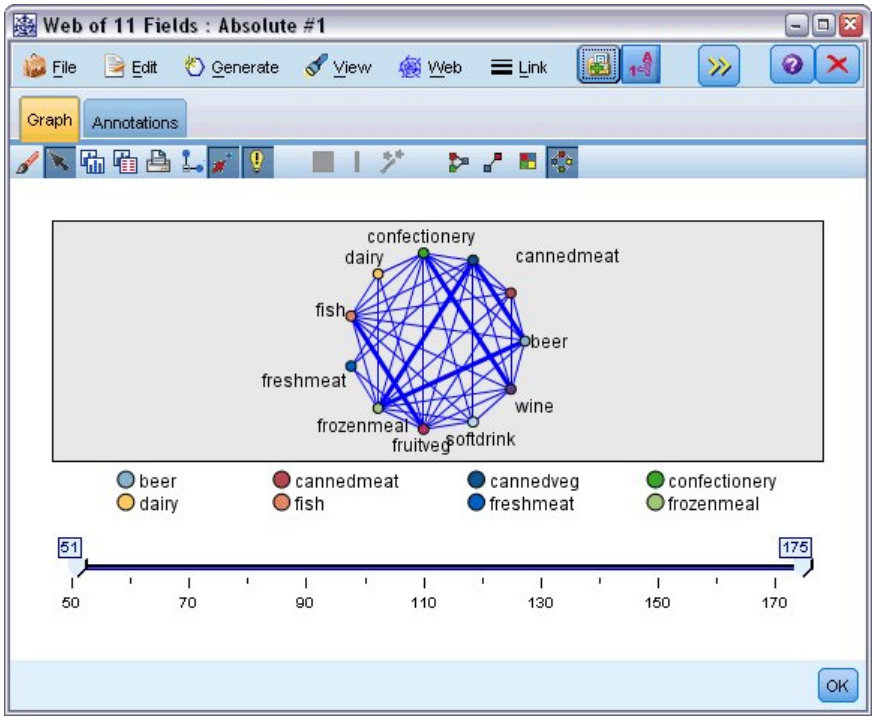


Figura 383. Exibição Web de associações de produtos

Como a maioria das combinações de categorias de produtos ocorre em várias cestas básicas, os fortes links nesta web são muito numerosos para mostrar os grupos de clientes sugeridos pelo modelo.

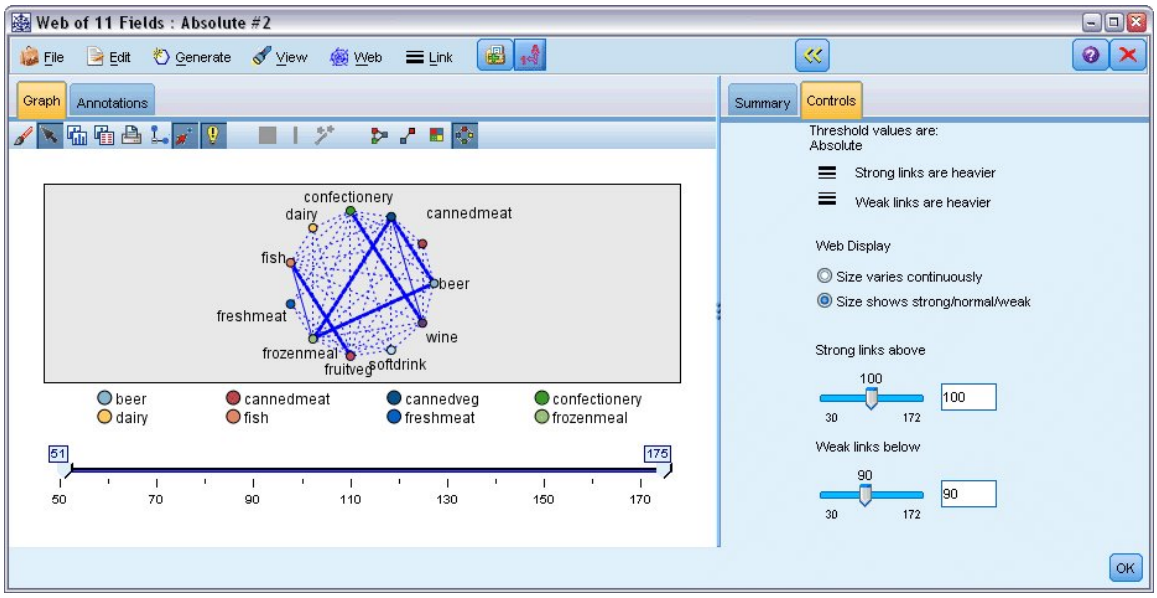


Figura 384. Exibição da web restrita

1. Para especificar conexões fracas e fortes, clique no botão de seta dupla amarela na barra de ferramentas. Isso expande a caixa de diálogo mostrando o resumo e controles de saída da web.
2. Selecione **Tamanho mostra redu/normal/fraco**.
3. Configure links fracos abaixo de 90.
4. Configurar links fortes acima de 100.

No display resultante, três grupos de clientes se destacam:

- Aqueles que compram peixe e frutas e legumes, que podem ser chamados de "comedores saudáveis"
- Aqueles que compram vinho e confeitaria
- Aqueles que compram cerveja, refeições congeladas, e legumes enlatados ("cerveja, feijão e pizza")

Perfil dos Grupos de Clientes

Você agora identificou três grupos de clientes com base nos tipos de produtos que eles compram, mas também gostaria de saber quem são esses clientes-ou seja, seu perfil demográfico. Isso pode ser alcançado com a identificação de cada cliente com uma bandeira para cada um desses grupos e usando a indução de regra (C5.0) para construir perfis baseados em regras dessas bandeiras.

Primeiro, deve-se derivar uma bandeira para cada grupo. Isso pode ser gerado automaticamente usando o display web que você acabou de criar. Usando o botão direito do mouse, clique no link entre *infrutveg* e *fish* para destacá-lo, em seguida, clique com o botão direito do mouse e selecione **Gerar Node do Link para o Link.**

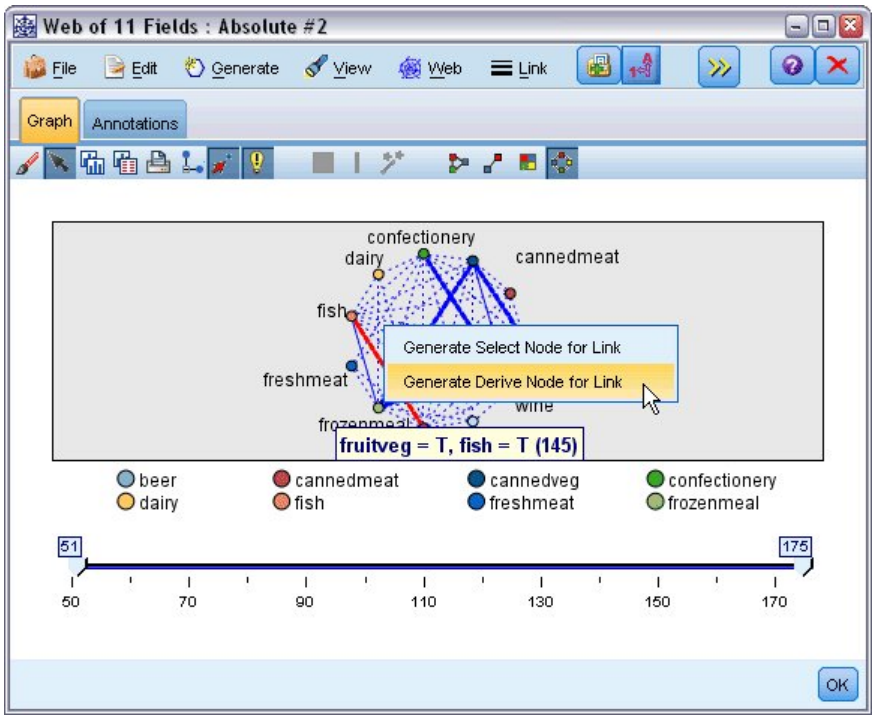


Figura 385. Derivando uma sinalização para cada grupo de clientes

Edite o nó de Derivação resultante para alterar o nome do campo Derivar para *saudáveis*. Repita o exercício com o link de *vinho* para *confectionery*, nomeando o campo resultante Derivativos *wine_chocs*.

Para o terceiro grupo (envolvendo três links), primeiro certise-se de que nenhum link é selecionado. Em seguida, selecione todos os três links no triângulo *cannedveg*, *beere frozenmeal* mantendo pressionada a tecla shift enquanto você clica no botão esquerdo do mouse. (Certifique-se de que você está no modo Interativo em vez de Editar modo.) Em seguida, a partir dos menus de exibição web escolha:

Gerar > **Derivar Nó ("E")**

Alterar o nome do campo resultante Derivar para *beer_beans_pizza*.

Para traçar o perfil desses grupos de clientes, conecte o nó do Tipo existente a estes três nós de Derivação em série e, em seguida, anexe outro nó Type. No novo nó Type, configure a função de todos os campos para *Nenhum*, exceto para *value*, *pmethod*, *sex*, *homeown*, *renda*, e *idade*, que deve ser configurado como *Entrada*, e o grupo de clientes relevante (por exemplo, *beer_beans_pizza*), que deve ser configurado como *Destino*. Conecte um nó C5.0 , configure o tipo de saída para **Conjunto de regras**e

clique em executar no nó. O modelo resultante (para *beer_beans_pizza*) contém um perfil demográfico claro para este grupo de clientes:

```
Rule 1 for T:
if sex = M
and income <= 16,900
then T
```

O mesmo método pode ser aplicado às outras bandeiras do grupo de clientes, selecionando-as como a saída no nó do segundo Tipo. Uma gama mais ampla de perfis alternativos pode ser gerada usando-se Apriori em vez de C5.0 neste contexto; Apriori também pode ser usado para traçar o perfil de todas as bandeiras do grupo de clientes simultaneamente porque ele não se restringe a um único campo de saída.

Resumo

Este exemplo revela como IBM SPSS Modelador pode ser usado para descobrir afinidades, ou links, em um banco de dados, tanto por modelagem (usando Apriori) quanto por visualização (usando um display web). Esses links correspondem a agrupamentos de casos nos dados, e esses grupos podem ser investigados em detalhes e proarquivados por modelagem (usando conjuntos de regras C5.0).

No domínio do varejo, tais agrupamentos de clientes podem, por exemplo, ser usados para destinar ofertas especiais para melhorar as taxas de resposta a mailings diretos ou para customizar a gama de produtos estocados por um ramo para adequar as demandas de sua base demográfica.

Capítulo 28. Avaliação De Novas Ofertas De Veículos (KNN)

Análise do Vizinho mais Próximo é um método de classificação de casos com base na sua similaridade com outros casos. Em aprendizado por máquina, ela foi desenvolvida como uma maneira de reconhecer padrões de dados sem requerer uma correspondência exata com nenhum dos padrões ou casos armazenados. Casos semelhantes ficam próximos uns dos outros e os casos diferentes ficam distantes uns dos outros. Portanto, a distância entre dois casos é uma medida de sua dissimilaridade.

Casos próximos são chamados de "vizinhos". Quando um novo caso (validação) é apresentado, sua distância de cada um dos casos no modelo é calculada. As classificações dos casos mais similares – os vizinhos mais próximos – são verificadas e o novo caso é colocado na categoria que contiver o maior número de vizinhos mais próximos.

É possível especificar o número de vizinhos mais próximos a serem examinados; este valor é denominado *k*. As fotos mostram como um novo caso seria classificado usando dois valores diferentes de *k*. Quando *k* = 5, o novo caso é colocado na categoria *1* porque a maioria dos vizinhos mais próximos pertence à categoria *1*. No entanto, quando *k* = 9, o novo caso é colocado na categoria *0* porque uma maioria dos vizinhos mais próximos pertence à categoria *0*.

A análise do vizinho mais próximo também pode ser utilizada para calcular valores para uma variável resposta contínua. Nesta situação, a média ou mediana do valor dos vizinhos mais próximos é utilizada para obter o valor predito para o novo caso.

Um fabricante de automóveis desenvolveu protótipos para dois novos veículos, um carro e um caminhão. Antes de introduzir os novos modelos em sua gama, a fabricante quer determinar quais os veículos existentes no mercado mais gostam dos protótipos-ou seja, quais os veículos são seus "vizinhos mais próximos" e, portanto, quais modelos eles estarão competindo contra.

A fabricante coletou dados sobre os modelos existentes sob uma série de categorias, e adicionou os detalhes de seus protótipos. As categorias sob as quais os modelos devem ser comparados incluem preço em milhares (*preço*), tamanho do motor (*engine_s*), horsepower (*horsepow*), entre-eixos (*wheelbas*), largura (*largura*), comprimento (*comprimento*), freio de peso (*curb_wgt*), capacidade de combustível (*fuel_cap*) e eficiência de combustível (*mpg*).

Este exemplo usa o fluxo denominado *car_sales_knn.str*, disponível na pasta *Demos* sob a subpasta *streams* O arquivo de dados é *car_sales_knn_mod.sav*.. Veja o tópico “Pasta Demos” na página 4 para obter mais informações.

Criando o Fluxo

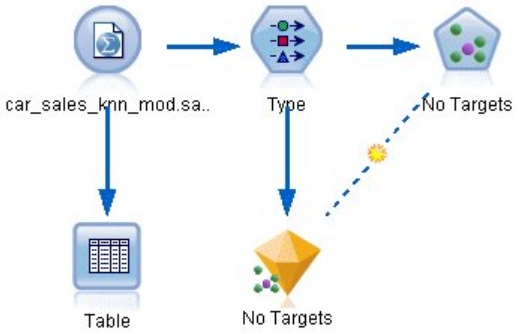
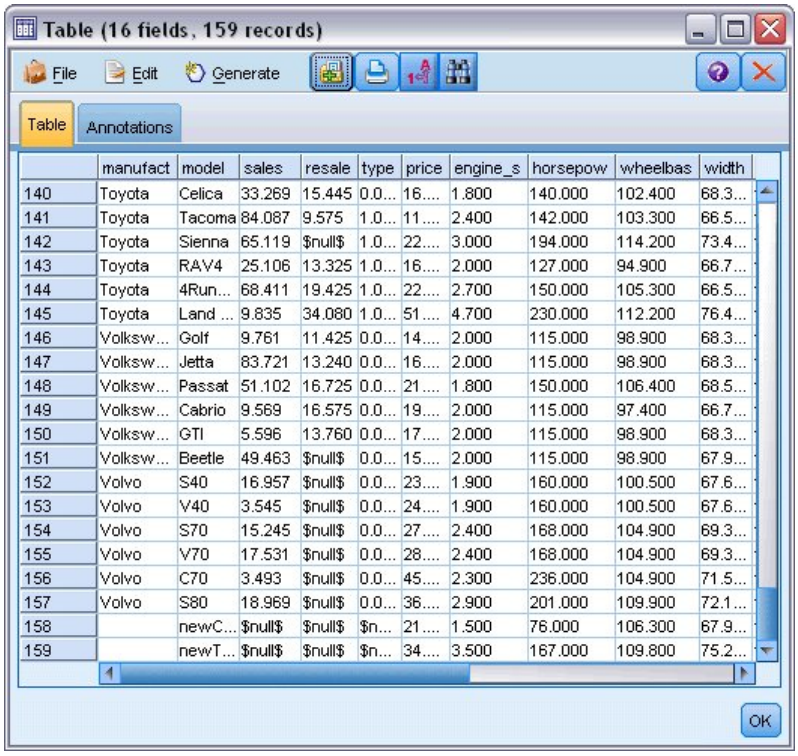


Figura 386. Fluxo de amostra para modelagem KNN

Crie um novo fluxo e inclua um nó de origem do Arquivo de Estatísticas apontando para *car_sales_knn_mod.sav* na pasta *Demos* de sua instalação do IBM SPSS Modelador .

Primeiro, vamos ver quais dados o fabricante coletou.

1. Conecte um nó da Tabela ao nó de origem do Arquivo de Estatísticas.
2. Abra o nó da Tabela e clique em **Executar**.



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

Figura 387. Dados de origem para carros e caminhões

Os detalhes dos dois protótipos, denominados *newCar* e *newTruck*, foram adicionados no final do arquivo.

Podemos ver a partir dos dados de origem que o fabricante está usando a classificação de "truck" (valor de 1 na coluna *tipo*) bastante vagamente para significar qualquer tipo de veículo não automóvel.

A última coluna, *partição*, é necessária a fim de que os dois protótipos possam ser designados como holdouts quando viemos identificar seus vizinhos mais próximos. Dessa forma, seus dados não influenciarão os cálculos, já que é o resto do mercado que queremos considerar. Configurando o valor *partição* dos dois registros de holdout para 1, enquanto todos os outros registros possuem um 0 neste campo, possibilita-nos usar este campo mais tarde quando chegamos a configurar os registros focais -- os registros para os quais queremos calcular os vizinhos mais próximos.

Deixe a janela de saída da tabela aberta por enquanto, como nos referiremos a ela mais tarde.

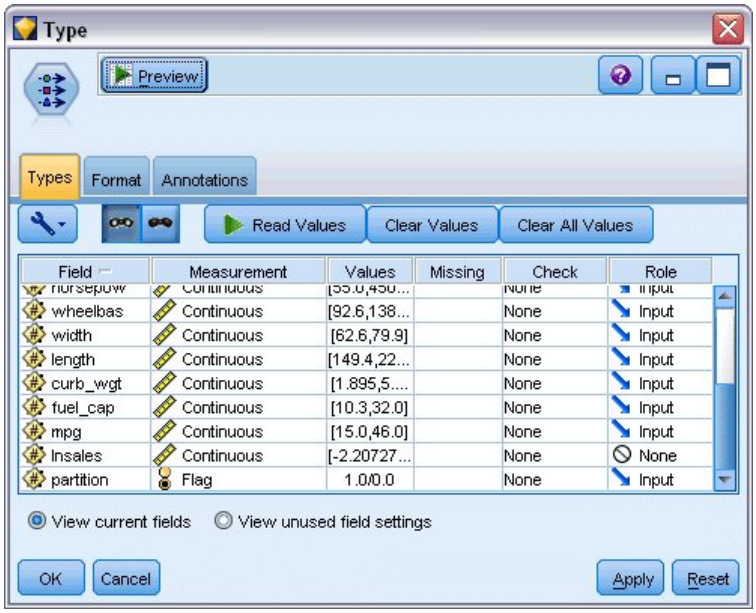


Figura 388. Configurações do nó do tipo

3. Inclua um nó Tipo no fluxo.
4. Conecte o nó Tipo ao nó de origem do Arquivo de Estatísticas.
5. Abra o nó Tipo.

   Queremos fazer a comparação apenas nos campos *preço* através do *mpg*, assim deixaremos a função para todos esses campos configurados para **Entrada**.
6. Configure a função para todos os outros campos (*manufact* através de *type*, mais *lnsales*) para **Nenhum**.
7. Configure o nível de medição para o último campo, *partição*, para **Flag**. Certise-se de que sua função esteja configurada como **Entrada**.
8. Clique em **Valores de leitura** para ler os valores de dados no fluxo.
9. Clique em **OK**.

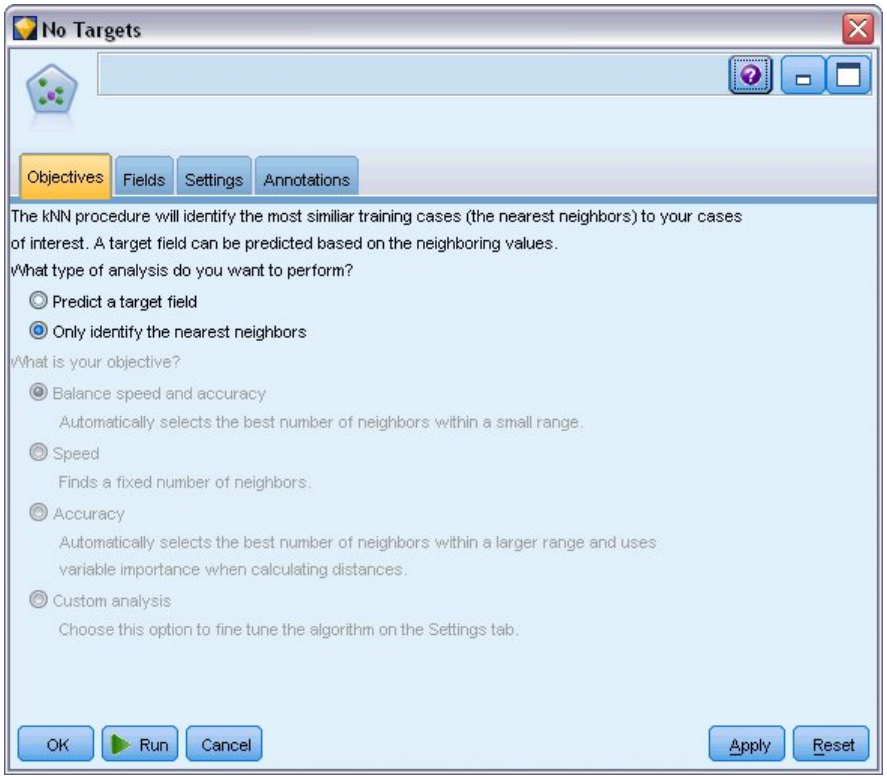


Figura 389. Optando por identificar os vizinhos mais próximos

10. Conecte um nó KNN ao nó Type.
11. Abra o nó KNN.

    Não vamos estar prevendo um campo alvo desta vez, porque só queremos encontrar os vizinhos mais próximos para os nossos dois protótipos.
12. Na guia **objetivos** , escolha **Apenas identificar os vizinhos mais próximos**.
13. Clique na guia **Configurações**.

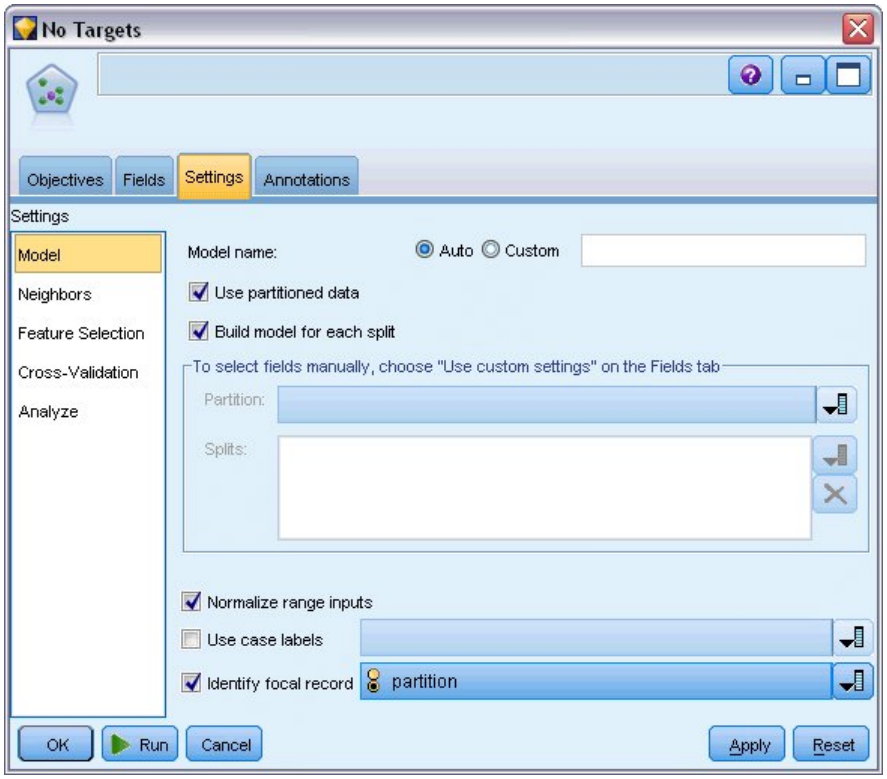


Figura 390. Como usar o campo de partição para identificar os registros focais

Agora podemos usar o campo *partição* para identificar os registros focais -- os registros para os quais queremos identificar os vizinhos mais próximos. Usando um campo de sinalização, nós asseguramos que registros onde o valor deste campo está configurado para 1 tornam-se nossos registros focais.

Como vimos, os únicos registros que têm o valor 1 nesse campo são *newCar* e *newTruck*, portanto, esses serão nossos registros focais.

14. No painel **Modelo** da guia **Configurações** , selecione a caixa de seleção **Identificar registro focal** .
15. A partir da lista suspensa para este campo, escolha **partição**.
16. Clique no botão **Executar**.

Examinando a saída

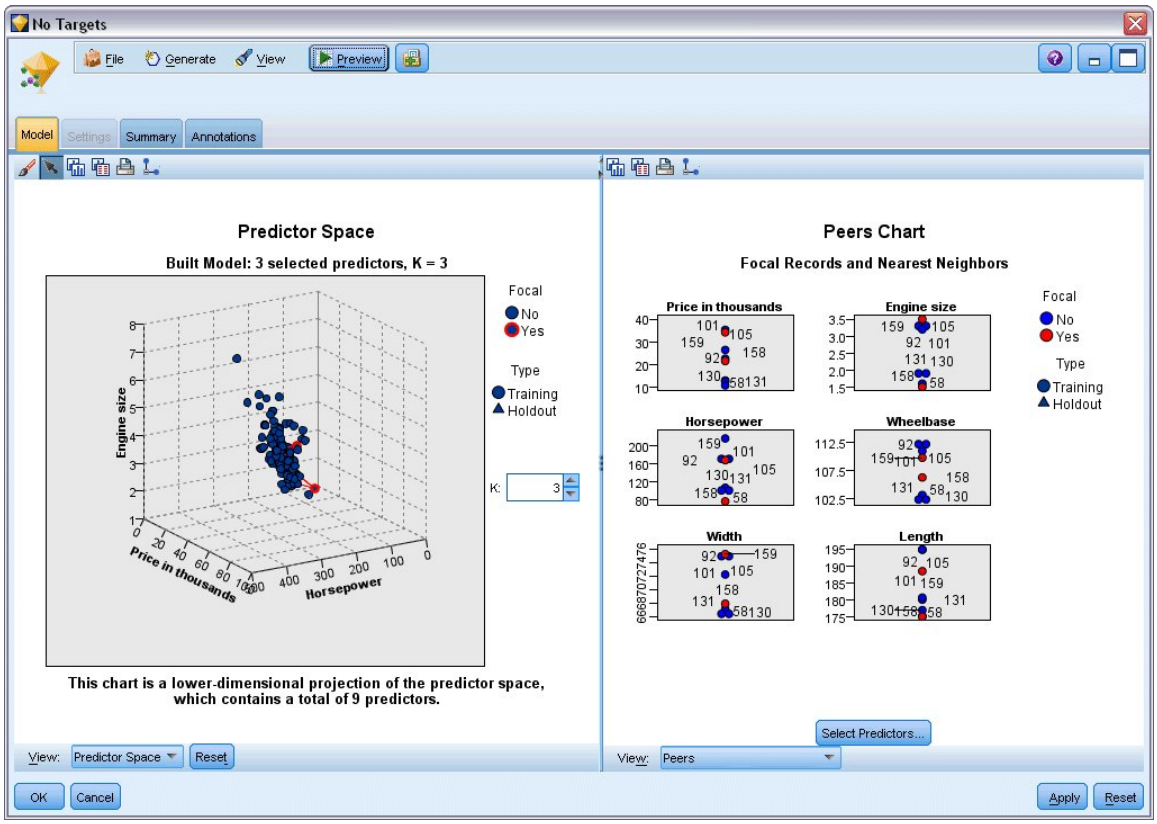


Figura 391. A janela do Model Viewer

Um nugget modelo foi criado na tela do fluxo e na paleta de Modelos. Abra qualquer um dos nuggets para ver o display Model Viewer, que possui uma janela de dois painéis:

- O primeiro painel exibe uma visão geral do modelo chamada de visualização principal. A visualização principal para o modelo Vizinho Mais Próximo é conhecida como o **espaço do preditor**.
- O segundo painel exibe um dos dois tipos de visualizações:

  Uma visualização de modelo auxiliar mostra mais informações sobre o modelo, mas não é focada no modelo em si.

  Uma visualização vinculada é uma visão que mostra detalhes sobre uma característica do modelo quando você perfuma em parte da visão principal.

Espaço de preditor

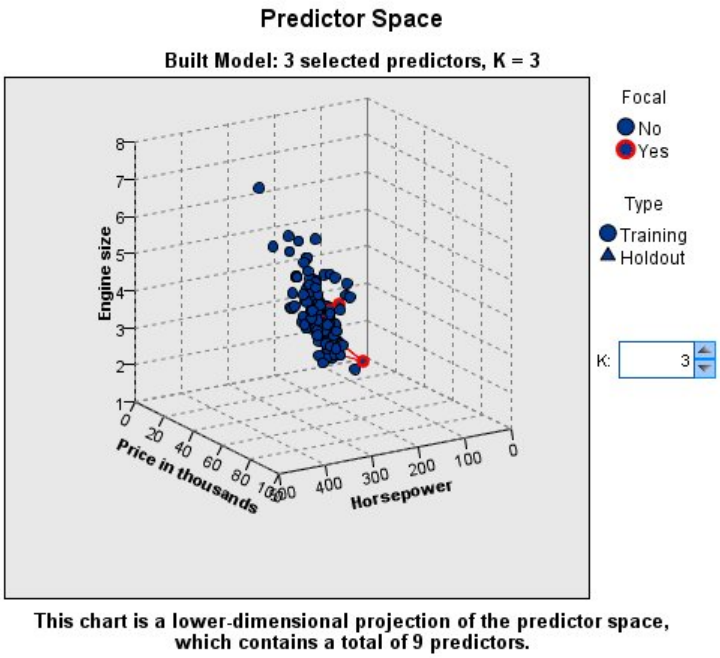


Figura 392. gráfico de espaço do preditor

O gráfico de espaço do preditor é um gráfico 3-D interativo que plota pontos de dados para três recursos (na verdade os três primeiros campos de entrada dos dados de origem), representando preço, tamanho do motor e potência.

Nossos dois registros focais são destacados em vermelho, com linhas conectando-os ao seu *k* vizinhos mais próximos.

Ao clicar e arrastar o gráfico, é possível girá-lo para obter uma melhor visualização da distribuição de pontos no espaço do preditor. Clique no botão **Reset** para devolvê-lo à visualização padrão.

Gráfico de Pares

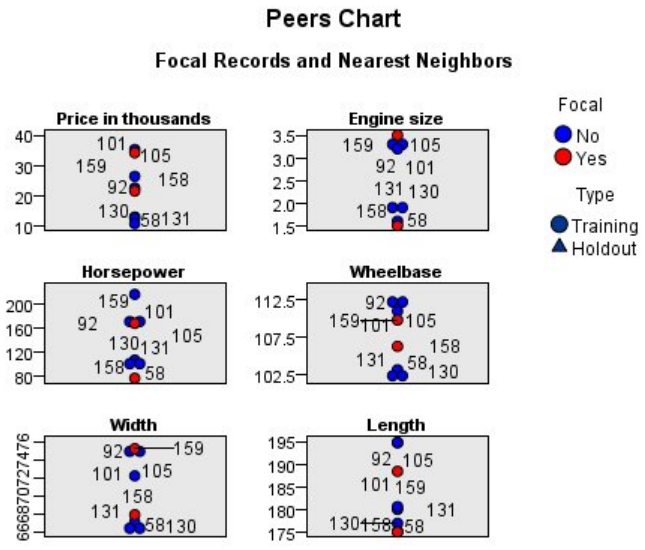


Figura 393. Gráfico de Peers

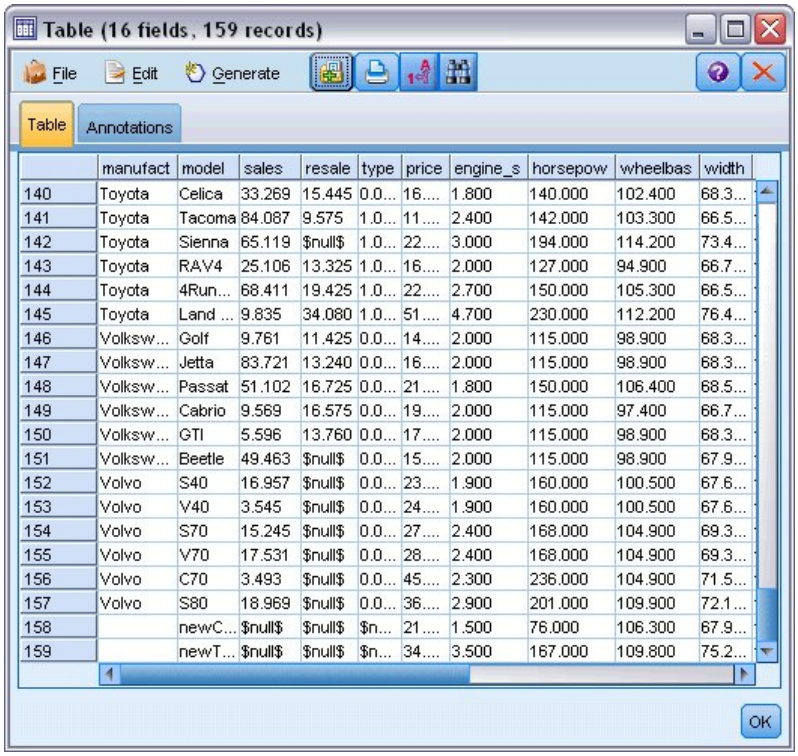
A visualização auxiliar padrão é o gráfico de pares, que destaca os dois registros focais selecionados no espaço do preditor e seus *k* vizinhos mais próximos em cada um dos seis recursos -- os seis primeiros campos de entrada dos dados de origem.

Os veículos são representados por seus números de registro nos dados de origem. É aqui que precisamos da saída do nó da Tabela para ajudar a identificá-los.

Se a saída do nó da Tabela ainda estiver disponível:

1. Clique na guia **Outputs** da pane manager na parte superior direita da janela principal IBM SPSS Modelador .
2. Clique duas vezes na entrada **Tabela (16 campos, 159 registros)**.

   Se a saída de tabela não estiver mais disponível:
3. Na janela principal IBM SPSS Modelador , abra o nó da Tabela.
4. Clique em **Executar** .



	manufact	model	sales	resale	type	price	engine_s	horsepow	wheelbas	width
140	Toyota	Celica	33.269	15.445	0.0...	16....	1.800	140.000	102.400	68.3...
141	Toyota	Tacoma	84.087	9.575	1.0...	11....	2.400	142.000	103.300	66.5...
142	Toyota	Sienna	65.119	$null$	1.0...	22....	3.000	194.000	114.200	73.4...
143	Toyota	RAV4	25.106	13.325	1.0...	16....	2.000	127.000	94.900	66.7...
144	Toyota	4Run...	68.411	19.425	1.0...	22....	2.700	150.000	105.300	66.5...
145	Toyota	Land ...	9.835	34.080	1.0...	51....	4.700	230.000	112.200	76.4...
146	Volksw...	Golf	9.761	11.425	0.0...	14....	2.000	115.000	98.900	68.3...
147	Volksw...	Jetta	83.721	13.240	0.0...	16....	2.000	115.000	98.900	68.3...
148	Volksw...	Passat	51.102	16.725	0.0...	21....	1.800	150.000	106.400	68.5...
149	Volksw...	Cabrio	9.569	16.575	0.0...	19....	2.000	115.000	97.400	66.7...
150	Volksw...	GTI	5.596	13.760	0.0...	17....	2.000	115.000	98.900	68.3...
151	Volksw...	Beetle	49.463	$null$	0.0...	15....	2.000	115.000	98.900	67.9...
152	Volvo	S40	16.957	$null$	0.0...	23....	1.900	160.000	100.500	67.6...
153	Volvo	V40	3.545	$null$	0.0...	24....	1.900	160.000	100.500	67.6...
154	Volvo	S70	15.245	$null$	0.0...	27....	2.400	168.000	104.900	69.3...
155	Volvo	V70	17.531	$null$	0.0...	28....	2.400	168.000	104.900	69.3...
156	Volvo	C70	3.493	$null$	0.0...	45....	2.300	236.000	104.900	71.5...
157	Volvo	S80	18.969	$null$	0.0...	36....	2.900	201.000	109.900	72.1...
158		newC...	$null$	$null$	$n...	21....	1.500	76.000	106.300	67.9...
159		newT...	$null$	$null$	$n...	34....	3.500	167.000	109.800	75.2...

Figura 394. Identificando registros por número de registro

Rolando até a parte inferior da tabela, podemos ver que *newCar* e *newTruck* são os dois últimos registros nos dados, números 158 e 159, respectivamente.

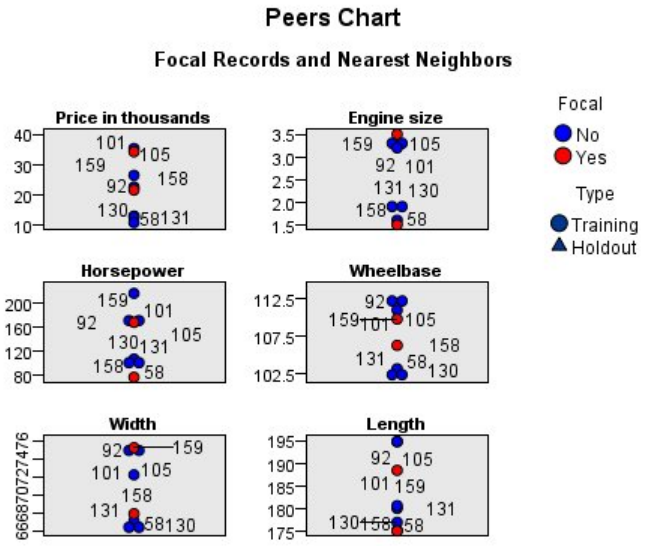


Figura 395. Como comparar recursos no gráfico de pares

A partir disso, podemos ver no gráfico de peers, por exemplo, que *newTruck* (159) possui um tamanho de mecanismo maior do que qualquer um de seus vizinhos mais próximos, enquanto *newCar* (158) possui um mecanismo menor do que qualquer *seus* vizinhos mais próximos.

Para cada um dos seis recursos, é possível mover o mouse sobre os pontos individuais para ver o valor real de cada recurso para aquele caso específico.

Mas quais veículos são os vizinhos mais próximos do *newCar* e do *newTruck*?

O gráfico de pares está um pouco lotado, então vamos mudar para uma visão mais simples.

5. Clique na lista suspensa **Visualizar** na parte inferior do gráfico peers (a entrada que atualmente diz **Peers**).
6. Selecione **Tabela de Vizinhança e Distância**.

Tabela de Vizinho e Distância

k Nearest Neighbors and Distances

Displayed for Initial Focal Records

Focal Record	Nearest Neighbors			Nearest Distar	
	1	2	3	1	2
158	131	130	58	0.979	0.990
159	105	92	101	0.580	0.634

Figura 396. Tabela de Vizinho e Distância

Isso é melhor. Agora podemos ver os três modelos aos quais cada um dos nossos dois protótipos estão mais próximos no mercado.

Para *newCar* (registro focal 158) eles são o Saturn SC (131), o Saturn SL (130) e o Honda Civic (58).

Não há grandes surpresas lá -- todos os três são carros de tamanho médio, portanto, *newCar* deve se encaixar bem, particularmente com sua excelente eficiência de combustível.

Para *newTruck* (registro focal 159), os vizinhos mais próximos são Nissan Quest (105), Mercury Villager (92) e Mercedes M-Class (101).

Como vimos anteriormente, estes não são necessariamente caminhões no sentido tradicional, mas simplesmente veículos que são classificados como não sendo automóveis. Olhando a saída do nó de

Tabela para seus vizinhos mais próximos, podemos ver que *newTruck* é relativamente caro, além de ser um dos mais pesados de seu tipo. No entanto, a eficiência de combustível é novamente melhor do que seus rivais mais próximos, portanto, isso deve contar a seu favor.

Resumo

Vimos como você pode usar a análise mais próxima do vizinho para comparar um amplo conjunto de recursos em casos a partir de um determinado conjunto de dados. Nós também calculamos, para dois registros de holdout muito diferentes, os casos que mais se assemelham a esses holdouts.

Capítulo 29. Descobrindo relacionamentos causais nas métricas de negócios (TCM)

Um negócio acompanha vários indicadores de desempenho chave que descrevem o estado financeiro do negócio ao longo do tempo, e também acompanham diversas métricas que podem controlar. Eles estão interessados em utilizar a modelagem causal temporal para descobrir relações causais entre as métricas controláveis e os principais indicadores de desempenho. Eles também gostariam de saber sobre quaisquer relações causais entre os principais indicadores de desempenho.

O arquivo de dados `tcm_kpi.sav` contém dados semanais sobre os principais indicadores de desempenho e as métricas controláveis. Os dados para os principais indicadores de desempenho são armazenados em campos com o prefixo *KPI*. Os dados para as métricas controláveis são armazenados em campos com o prefixo *Lever*.

Criando o fluxo

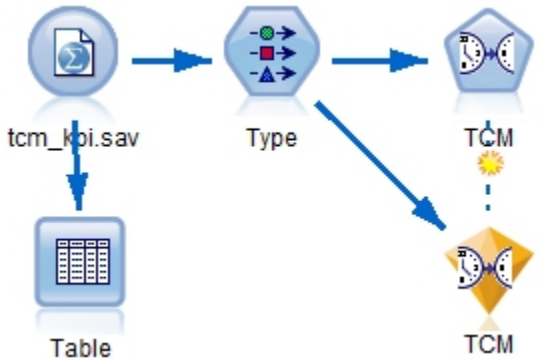
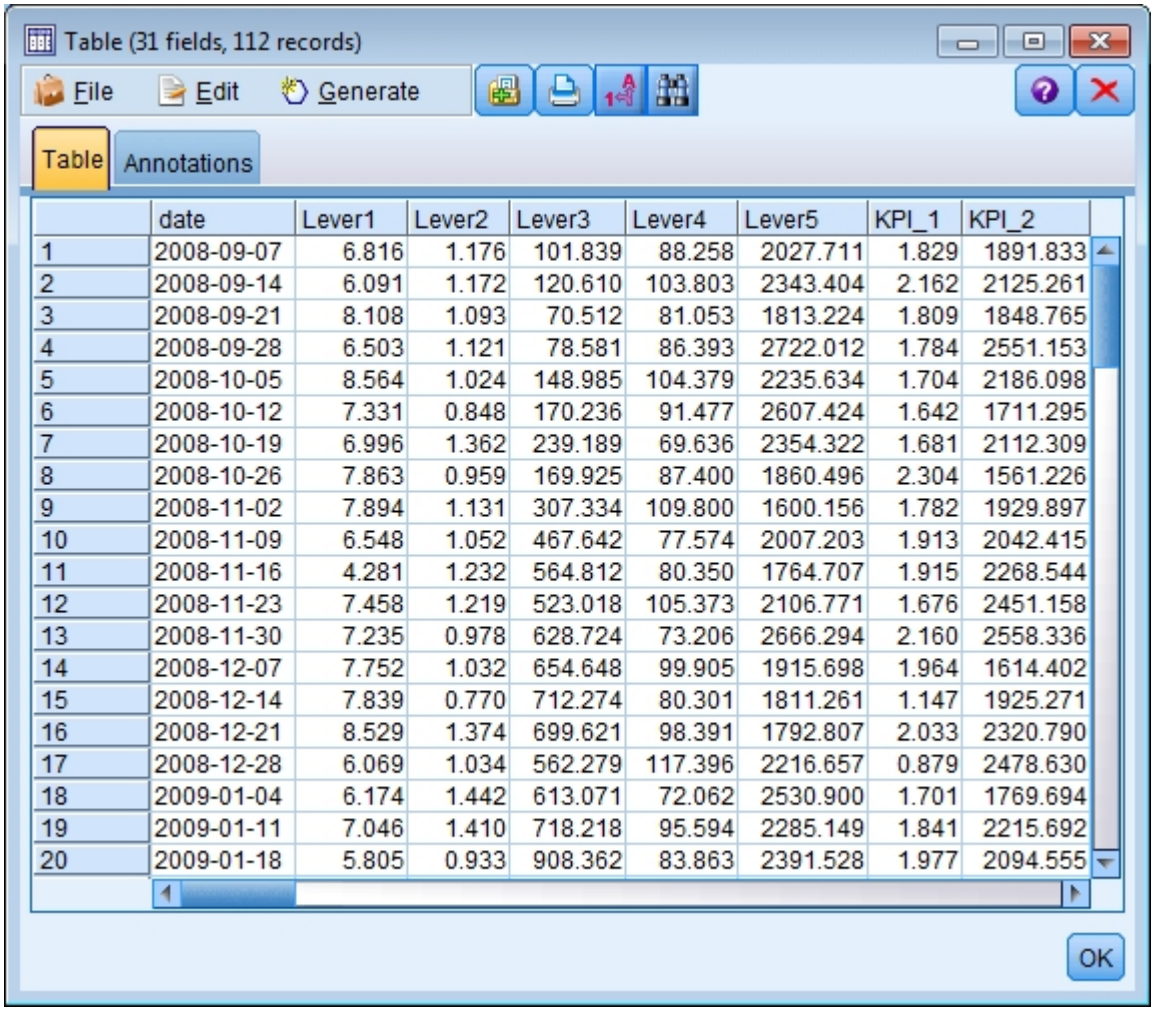


Figura 397. Fluxo de amostra para modelagem TCM

1. Crie um novo fluxo e inclua um nó de origem do Arquivo de Estatísticas apontando para *tcm_kpi.sav* na pasta *Demos* de sua instalação do IBM SPSS Modelador .
2. Conecte um nó da Tabela ao nó de origem do Arquivo de Estatísticas.
3. Abra o nó da Tabela e clique em **Executar** para dar uma olhada nos dados. Ele contém dados semanais sobre os principais indicadores de desempenho e as métricas controláveis. Os dados para os principais indicadores de desempenho são armazenados em campos com o prefixo *KPI*, e os dados para as métricas controláveis são armazenados em campos com o prefixo *Lever*.



	date	Lever1	Lever2	Lever3	Lever4	Lever5	KPI_1	KPI_2
1	2008-09-07	6.816	1.176	101.839	88.258	2027.711	1.829	1891.833
2	2008-09-14	6.091	1.172	120.610	103.803	2343.404	2.162	2125.261
3	2008-09-21	8.108	1.093	70.512	81.053	1813.224	1.809	1848.765
4	2008-09-28	6.503	1.121	78.581	86.393	2722.012	1.784	2551.153
5	2008-10-05	8.564	1.024	148.985	104.379	2235.634	1.704	2186.098
6	2008-10-12	7.331	0.848	170.236	91.477	2607.424	1.642	1711.295
7	2008-10-19	6.996	1.362	239.189	69.636	2354.322	1.681	2112.309
8	2008-10-26	7.863	0.959	169.925	87.400	1860.496	2.304	1561.226
9	2008-11-02	7.894	1.131	307.334	109.800	1600.156	1.782	1929.897
10	2008-11-09	6.548	1.052	467.642	77.574	2007.203	1.913	2042.415
11	2008-11-16	4.281	1.232	564.812	80.350	1764.707	1.915	2268.544
12	2008-11-23	7.458	1.219	523.018	105.373	2106.771	1.676	2451.158
13	2008-11-30	7.235	0.978	628.724	73.206	2666.294	2.160	2558.336
14	2008-12-07	7.752	1.032	654.648	99.905	1915.698	1.964	1614.402
15	2008-12-14	7.839	0.770	712.274	80.301	1811.261	1.147	1925.271
16	2008-12-21	8.529	1.374	699.621	98.391	1792.807	2.033	2320.790
17	2008-12-28	6.069	1.034	562.279	117.396	2216.657	0.879	2478.630
18	2009-01-04	6.174	1.442	613.071	72.062	2530.900	1.701	1769.694
19	2009-01-11	7.046	1.410	718.218	95.594	2285.149	1.841	2215.692
20	2009-01-18	5.805	0.933	908.362	83.863	2391.528	1.977	2094.555

Figura 398. Dados de origem para principais indicadores de desempenho e métricas controláveis

4. Inclua um nó Tipo no fluxo.
5. Conecte o nó Tipo ao nó de origem do Arquivo de Estatísticas.

executando a análise

1. Anexar um nó TCM para o nó Type, depois abra o nó TCM e vá até a seção **Observações** da guia **Campos** .

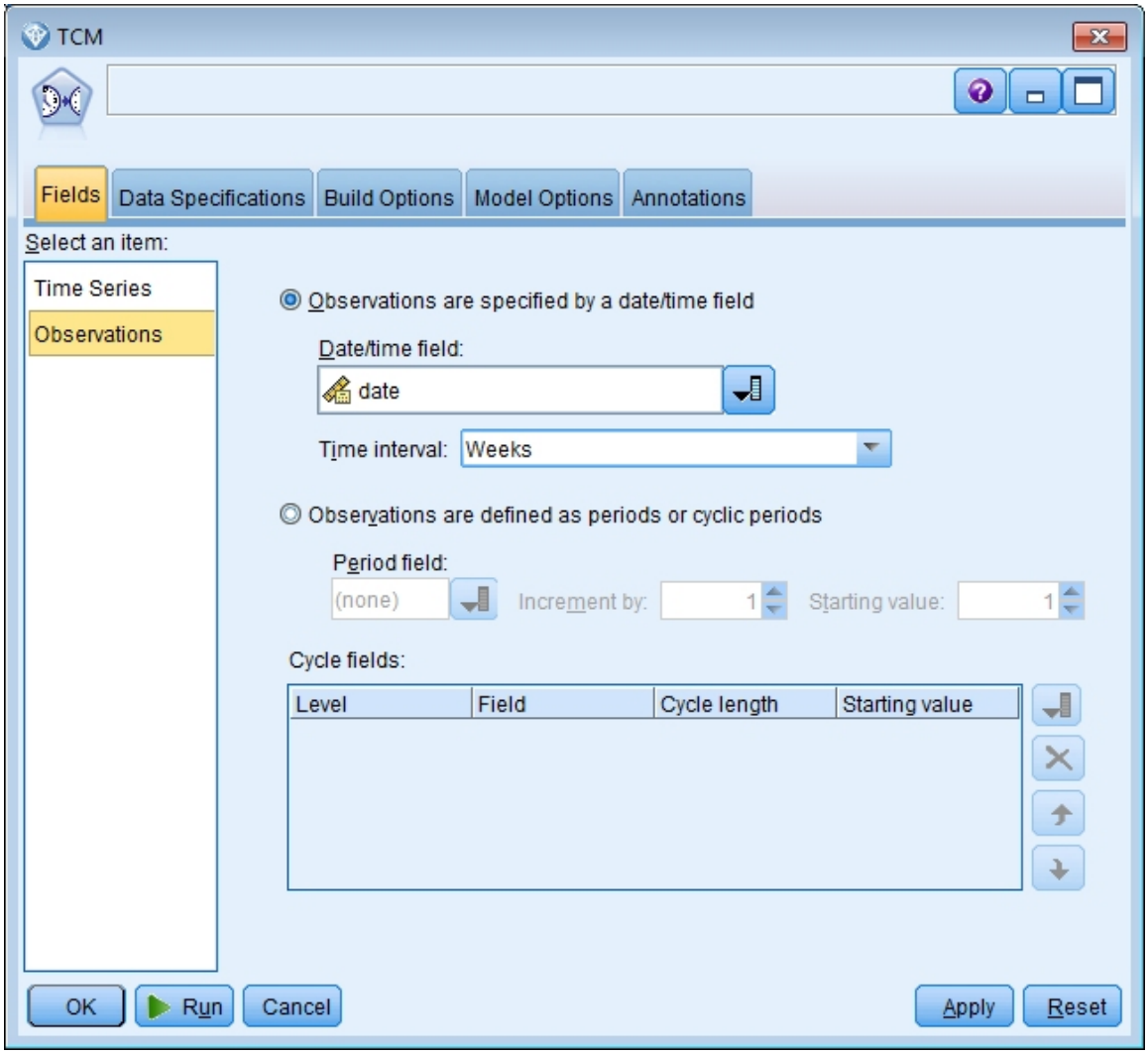


Figura 399. Modelagem Causal Temporal, observações

2. Selecione *date* a partir do campo Data / hora e selecione *Semanas* a partir do campo Intervalo de Tempo.
3. Clique em **Série Tempo** e selecione **Usar funções predefinidas**.

   No conjunto de dados de amostra `tcm_kpi.sav`, os campos *Lever1* por meio de *Lever5* possuem a função de Entrada e *KPI_1* através de *KPI_25* tem o papel de Ambos. Quando **Usar funções predefinidas** é selecionado, campos com uma função de Entrada são tratados como entradas e campos de candidatos com uma função de Ambos são tratados como ambos entradas e destinos de modelagem para modelagem causal temporal.

   O procedimento de modelagem causal temporal determina as melhores entradas para cada destino a partir do conjunto de entradas do candidato. Neste exemplo, as entradas do candidato são os campos *Lever1* através de *Lever5* e os campos *KPI_1* através de *KPI_25*.
4. Clique em **Executar** .

Gráfico de qualidade geral do modelo

O item de saída Qualidade do Modelo Geral, que é gerado por padrão, exibe um gráfico de barras e um traçado de ponto associado do modelo fit para todos os modelos. Há um modelo separado para cada série alvo. O modelo fit é medido pela estatística apta escolhida. Este exemplo usa a estatística de ajuste padrão, que é a R Square.

O item Qualidade do Modelo Geral contém recursos interativos. Para ativar os recursos, ative o item clicando duas vezes no gráfico de Qualidade do Modelo Geral no Viewer.

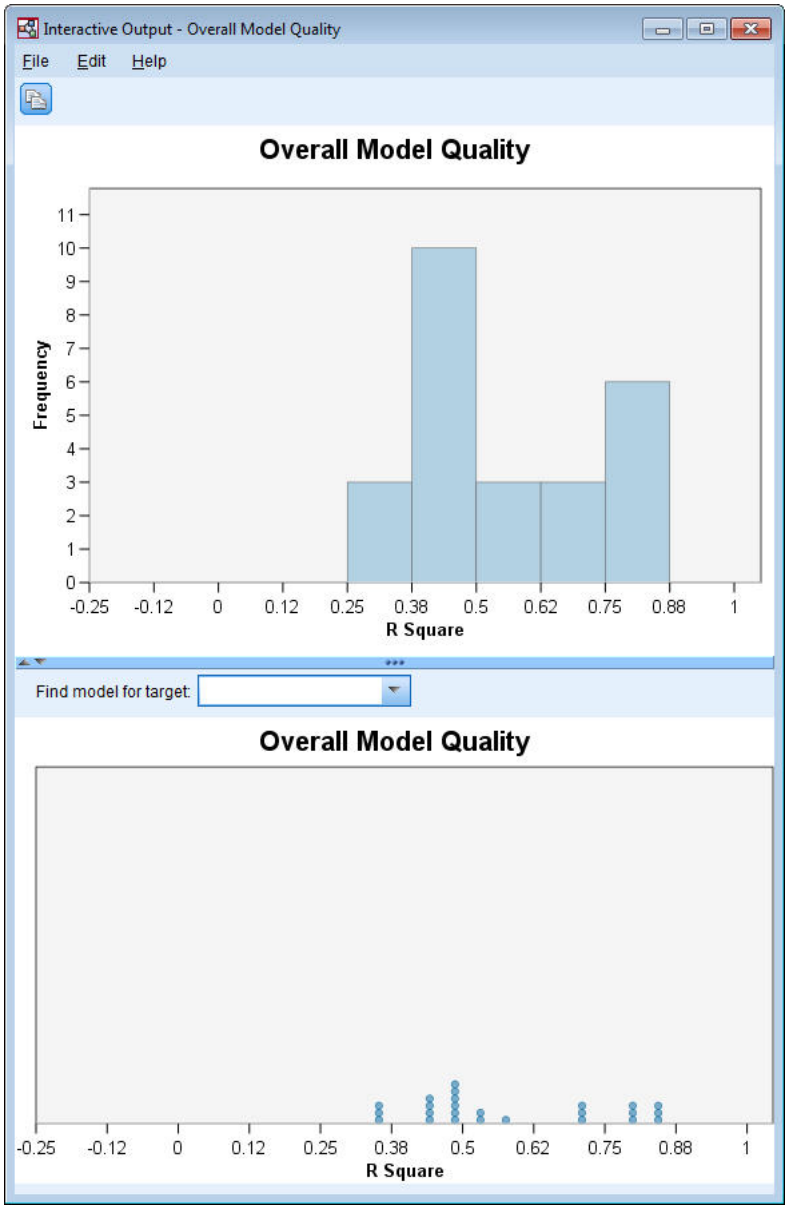


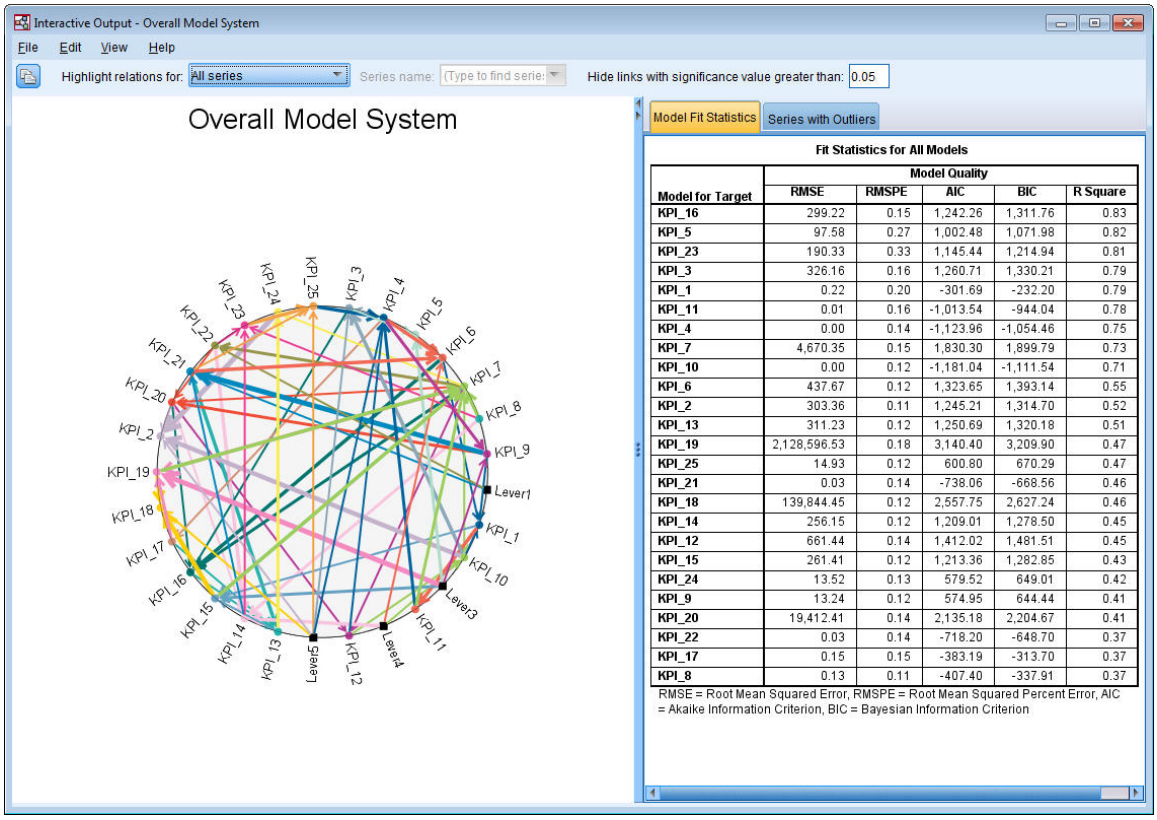
Figura 400. Qualidade de modelo geral

Clicar em uma barra no gráfico de barras filtra o gráfico de pontos, de modo que ele exiba apenas os modelos que estiverem associados com a barra selecionada. Pairar sobre um ponto no gráfico de pontos exibe uma dica de ferramenta que contém o nome da série associada e o valor da estatística fit. Você pode encontrar o modelo para uma determinada série de destino no gráfico de pontos especificando o nome da série na caixa **Localize modelo para destino** .

Sistema de modelo global

O item de saída do Sistema Modelo Geral, que é gerado por padrão, exibe uma representação gráfica das relações causais entre séries no sistema modelo. Por padrão, as relações para os 10 principais modelos são mostradas, conforme determinado pelo valor da estatística de ajuste da Praça R. O número de modelos top (também chamados de melhores modelos de encaixe) e a estatística fit são especificados nas configurações da Série a Exibir (na guia Opções de Construção) do diálogo Temporal Causal Modelagem.

O item Sistema Modelo Geral contém recursos interativos. Para ativar os recursos, ative o item clicando duas vezes no gráfico do Sistema Modelo Geral no Viewer. Neste exemplo, é mais importante ver as relações entre todas as séries do sistema. Na saída interativa, selecione **Todas as séries** a partir da lista suspensa **Relações Highlight para** .



Model for Target	Model Quality				
	RMSE	RMSPE	AIC	BIC	R Square
KPI_16	299.22	0.15	1,242.26	1,311.76	0.83
KPI_5	97.58	0.27	1,002.48	1,071.98	0.82
KPI_23	190.33	0.33	1,145.44	1,214.94	0.81
KPI_3	326.16	0.16	1,260.71	1,330.21	0.79
KPI_1	0.22	0.20	-301.69	-232.20	0.79
KPI_11	0.01	0.16	-1,013.54	-944.04	0.78
KPI_4	0.00	0.14	-1,123.96	-1,054.46	0.75
KPI_7	4,670.35	0.15	1,830.30	1,899.79	0.73
KPI_10	0.00	0.12	-1,181.04	-1,111.54	0.71
KPI_6	437.67	0.12	1,323.65	1,393.14	0.55
KPI_2	303.36	0.11	1,245.21	1,314.70	0.52
KPI_13	311.23	0.12	1,250.69	1,320.18	0.51
KPI_19	2,128,596.53	0.18	3,140.40	3,209.90	0.47
KPI_25	14.93	0.12	600.80	670.29	0.47
KPI_21	0.03	0.14	-738.06	-668.56	0.46
KPI_18	139,844.45	0.12	2,557.75	2,627.24	0.46
KPI_14	256.15	0.12	1,209.01	1,278.50	0.45
KPI_12	661.44	0.14	1,412.02	1,481.51	0.45
KPI_15	261.41	0.12	1,213.36	1,282.85	0.43
KPI_24	13.52	0.13	579.52	649.01	0.42
KPI_9	13.24	0.12	574.95	644.44	0.41
KPI_20	19,412.41	0.14	2,135.18	2,204.67	0.41
KPI_22	0.03	0.14	-718.20	-648.70	0.37
KPI_17	0.15	0.15	-383.19	-313.70	0.37
KPI_8	0.13	0.11	-407.40	-337.91	0.37

Figura 401. Sistema de Modelo Geral, visualização para todas as séries

Todas as linhas que conectam um determinado destino a suas entradas possuem a mesma cor, e a seta em cada linha aponta de uma entrada para o destino dessa entrada. Por exemplo, *Lever3* é uma entrada para *KPI_19*.

A espessura de cada linha indica o significado da relação causal, em que as linhas mais grossas representam uma relação mais significativa. Por padrão, relações causais com um valor de significância maior que 0.05 são ocultas. No nível 0.05 , apenas *Lever1*, *Lever3*, *Lever4*e *Lever5* têm relações causais significativas com os campos do principal indicador de desempenho. Você pode alterar o nível de significância de limite ao entrar em um valor no campo que é rotulado **Hide links com valor de significância maior que**.

Além de descobrir as relações causais entre os campos *Lever* e campos indicadores-chave de desempenho, a análise também descobria as relações entre os principais campos indicadores de desempenho. Por exemplo, *KPI_10* foi selecionado como uma entrada para o modelo para *KPI_2*.

Você pode filtrar a visualização para mostrar apenas as relações para uma única série. Por exemplo, para visualizar apenas as relações para *KPI_19*, clique na etiqueta para *KPI_19*, clique com o botão direito do mouse e selecione **Highlight relations for series**.

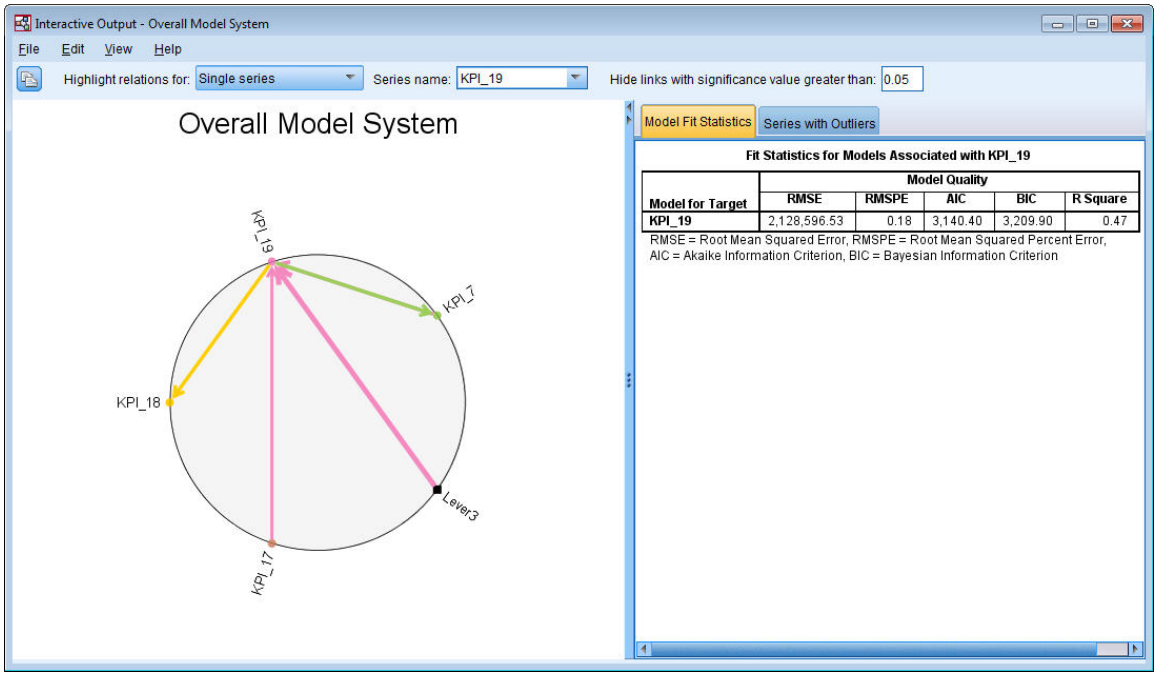


Figura 402. Sistema de Modelo Geral, visualização para série única

Esta visualização mostra as entradas para *KPI_19* que possuem um valor de significância menor ou igual a 0.05. Ele também mostra que, no nível de significância 0.05 , *KPI_19* foi selecionado como uma entrada para *KPI_18* e *KPI_7*.

Além de exibir as relações para a série selecionada, o item de saída também contém informações sobre eventuais outliers que foram detectados para a série. Clique na guia **Série com Outliers** .

Series with Outliers for KPI_19

Series	Time	Observed Value
KPI_19	2008-10-12	7,358,201.68
	2009-04-05	2.10E+007
	2010-09-19	6,492,157.97

Figura 403. Outliers para KPI_19

Três outliers foram detectados para *KPI_19*. Dado o sistema modelo, que contém todas as conexões descobertas, é possível ir além da detecção de outlier e determinar a série que mais provavelmente causa um determinado outlier. Este tipo de análise é referido como análise de causa raiz outlier e é coberto em um tópico posterior neste estudo de caso.

Diagramas de impacto

Você pode obter uma visão completa de todas as relações que estão associadas a uma determinada série gerando um diagrama de impacto. Clique na etiqueta para *KPI_19* no gráfico do Sistema Modelo Geral, clique com o botão direito do mouse e selecione **Criar Diagrama de Impacto**.

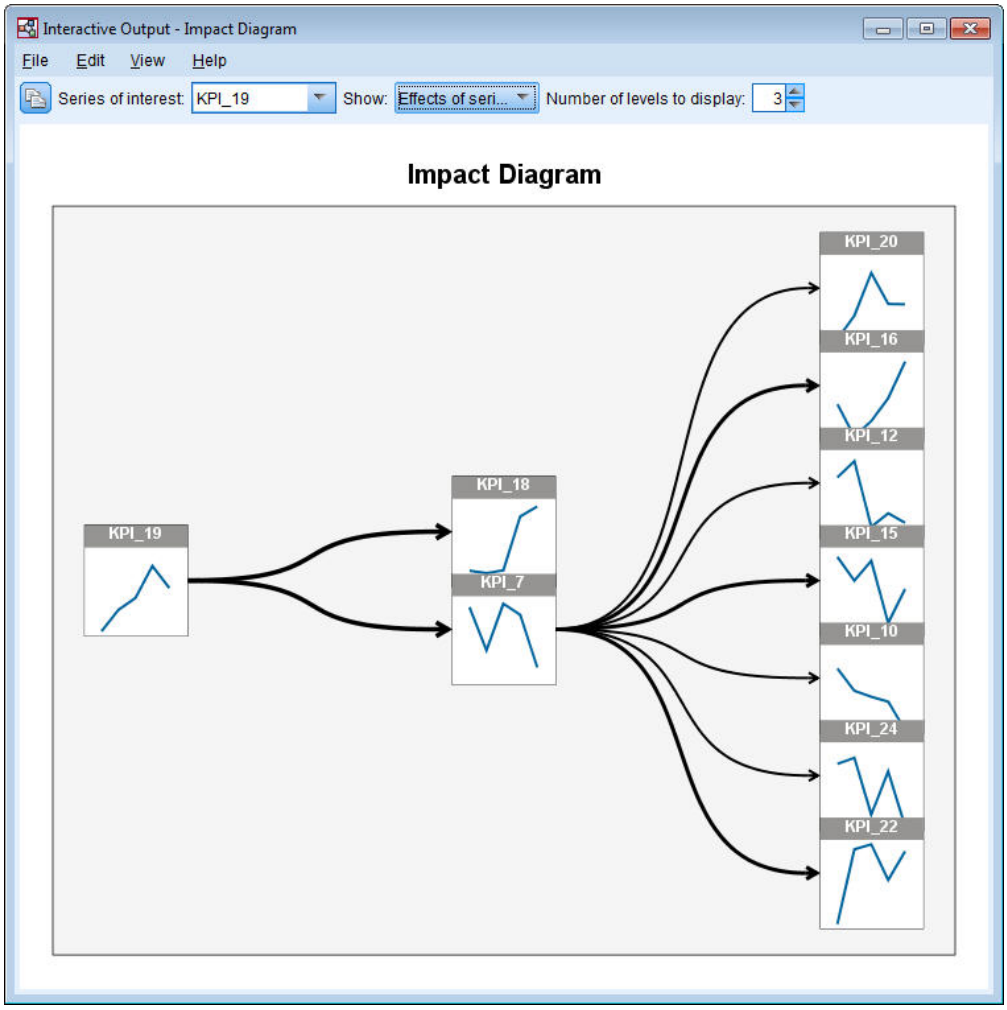


Figura 404. Diagrama De Impacto de efeitos

Quando um diagrama de impacto é criado a partir do Sistema Modelo Geral, como neste exemplo, ele mostra inicialmente a série que são afetadas pela série selecionada. Por padrão, os diagramas de impacto mostram três níveis de efeitos, em que o primeiro nível é apenas a série de interesse. Cada nível adicional mostra mais efeitos indiretos da série de interesse. Você pode alterar o valor do **Número de níveis a serem exibidos** para mostrar mais ou menos níveis de efeitos. O diagrama de impacto para este exemplo mostra que *KPI_19* é uma entrada direta para ambos *KPI_18* e *KPI_7*, mas ela afeta indiretamente uma série de séries através de seu efeito na série *KPI_7*. Como no sistema de modelo geral, a espessura das linhas indica o significado das relações causais.

O gráfico que é exibido em cada nó do diagrama de impacto mostra os últimos valores L+1 da série associada ao final do período de estimação e quaisquer valores previstos, em que L é o número de termos lag que estão incluídos em cada modelo. Você pode obter um gráfico de sequência detalhado desses valores através de um único clique no nó associado.

O duplo clique de um nó configura a série associada como a série de interesse, e regenera o diagrama de impacto baseado nessa série. Você também pode especificar um nome de série na caixa **Série de interesse** para selecionar uma série diferente de interesse.

Os diagramas de impacto também podem mostrar a série que afeta a série de interesse. Essas séries são referidas como *causas*. Para ver a série que afeta *KPI_19*, selecione **Causas da série** a partir do drop-down **Show** .

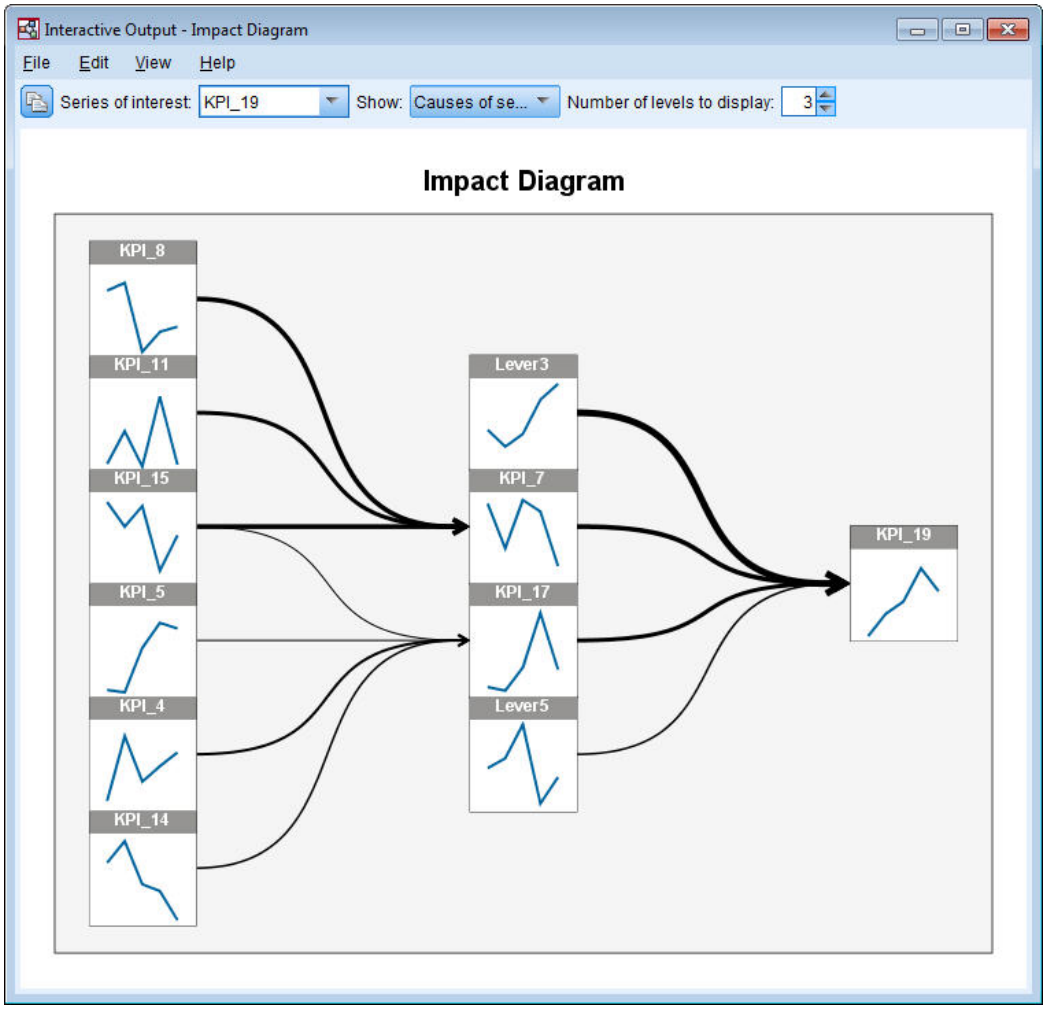


Figura 405. Diagrama De Impacto de causas

Esta visão mostra que o modelo para *KPI_19* tem quatro entradas e que *Lever3* tem a conexão causal mais significativa com *KPI_19*. Ele também mostra séries que afetam indiretamente *KPI_19* através de seus efeitos em *KPI_7* e *KPI_17*. O mesmo conceito de níveis que foi discutido para efeitos também se aplica às causas. Da mesma forma, você pode alterar o valor do **Número de níveis a serem exibidos** para mostrar mais ou menos níveis de causas.

Determinando causas raiz de outliers

Dado um sistema de modelo causal temporal, é possível ir além da detecção de outlier e determinar a série que mais provavelmente causa um determinado outlier. Esse processo é chamado de análise de causa raiz outlier e deve ser solicitado em série por base de série. A análise requer um sistema de modelo causal temporal e os dados que foram usados para construir o sistema. Neste exemplo, o dataset ativo são os dados que foram usados para construir o sistema modelo.

Para executar análise de causa raiz outlier:

1. No diálogo TCM, acesse a aba **Construir Opções** e, em seguida, clique em **Série a Exibir** na lista **Selecionar um item** .

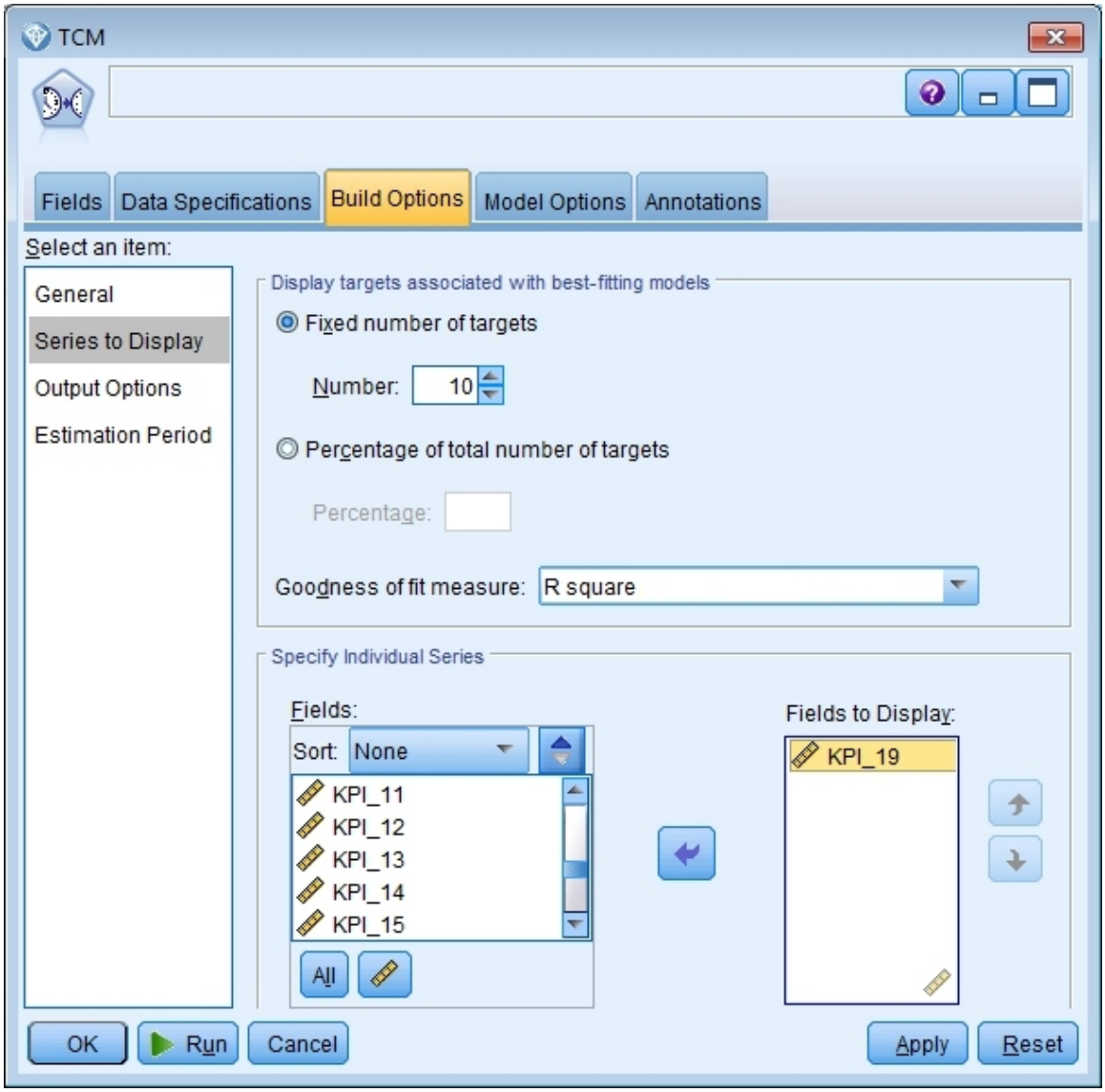


Figura 406. Temporal Causal Modelo Série a Exibição

2. Mova *KPI_19* para a lista **Campos a exibir** .
3. Clique em **Opções de saída** na lista **Selecionar um item** na guia Opções.

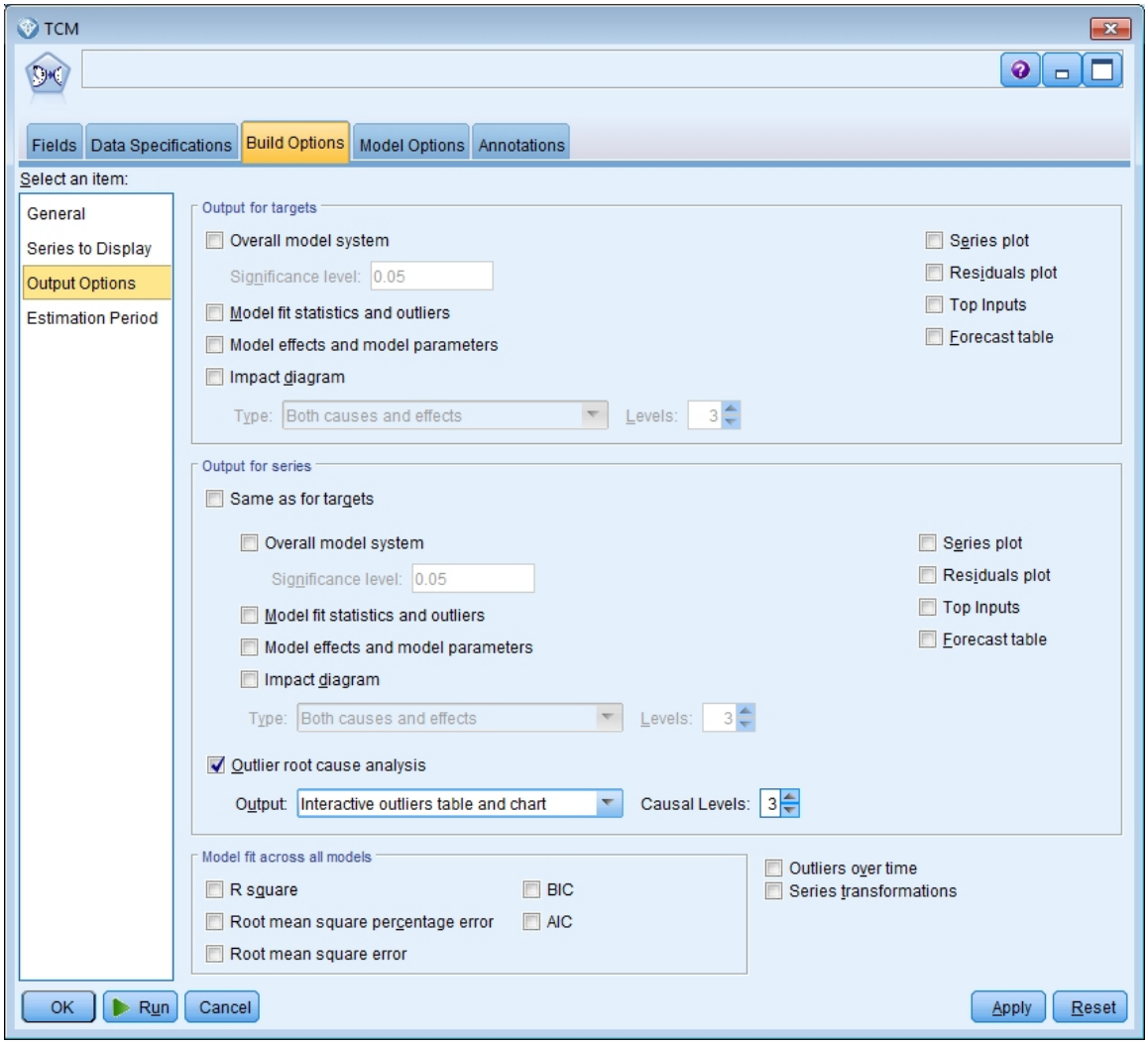


Figura 407. Opções de Saída de Modelo Causal Causal

4. Desmarque **Sistema de modelo geral**, **Mesmo que para destinos**, **R square**e **transformações da série**.
5. Selecione **Análise de causa raiz do Outlier** e mantenha as configurações existentes para **Output** e **Causal levels**.
6. Clique em **Executar** .
7. Clique duas vezes no gráfico de Análise de Raiz Externa do Outlier para *KPI_19* no Viewer para ativá-lo.

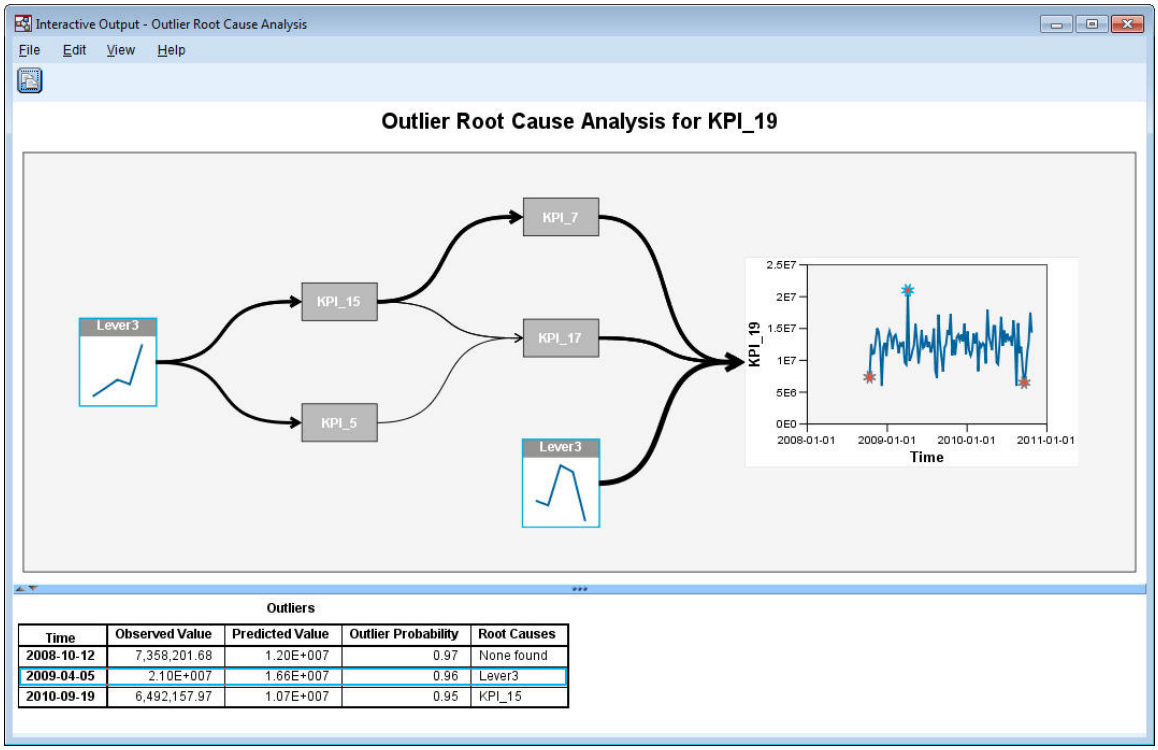


Figura 408. Análise De Causa Raiz Outlier para KPI_19

Os resultados da análise são resumidos na tabela Outliers. A tabela mostra que causas raiz foram encontradas para os outliers em 2009-04-05 e 2010-09-19, mas nenhuma causa raiz foi encontrada para o outlier em 2008-10-12. Clicar em uma linha na tabela Outliers destaca o caminho para a série de causa raiz, como mostrado aqui para o outlier em 2009-04-05. Esta ação também destaca o outlier selecionado no gráfico de sequência. Você também pode clicar no ícone para um outlier diretamente no gráfico de sequência para destacar o caminho para a série de causa raiz para aquele outlier.

Para o outlier em 2009-04-05, a causa raiz é *Lever3*. O diagrama mostra que *Lever3* é uma entrada direta para *KPI_19*, mas que também indiretamente influencia *KPI_19* através de seu efeito em outras séries que afetam *KPI_19*. Um dos parâmetros configuráveis para análise de causa raiz outlier é o número de níveis causais para pesquisar causas raiz. Por padrão, três níveis são pesquisados. As ocorrências das séries de causa raiz são exibidas até o número especificado de níveis causais. Neste exemplo, *Lever3* ocorre em ambos o primeiro nível causal e o terceiro nível causal.

Cada nó no caminho destacado para um outlier contém um gráfico cujo intervalo de tempo depende do nível no qual o nó ocorre. Para nós no primeiro nível causal, o intervalo é T-1 a T-L onde T é o tempo em que ocorre o outlier e L é o número de termos lag que estão incluídos em cada modelo. Para nós no segundo nível causal, o intervalo é T-2 a T-L-1; e para o terceiro nível o intervalo é T-3 a T-L-2. Você pode obter um gráfico de sequência detalhado desses valores através de um único clique no nó associado.

Cenários de Execução

Dado um sistema de modelo causal temporal, é possível executar cenários definidos pelo usuário. Um *cenário* é definido por uma série temporal, que é referida como a *série raiz* e um conjunto de valores definidos pelo usuário para essa série sobre um intervalo de tempo especificado. Os valores especificados são então usados para gerar predições para as séries temporais que são afetadas pela série raiz. A análise requer um sistema de modelo causal temporal e os dados que foram usados para construir o sistema. Neste exemplo, o dataset ativo são os dados que foram usados para construir o sistema modelo.

Para executar cenários:

1. No diálogo de saída TCM, clique no botão **Análise de Cenários** .
2. No diálogo Temporal Causal Model Cenários, clique em **Definir Período do Cenário**.

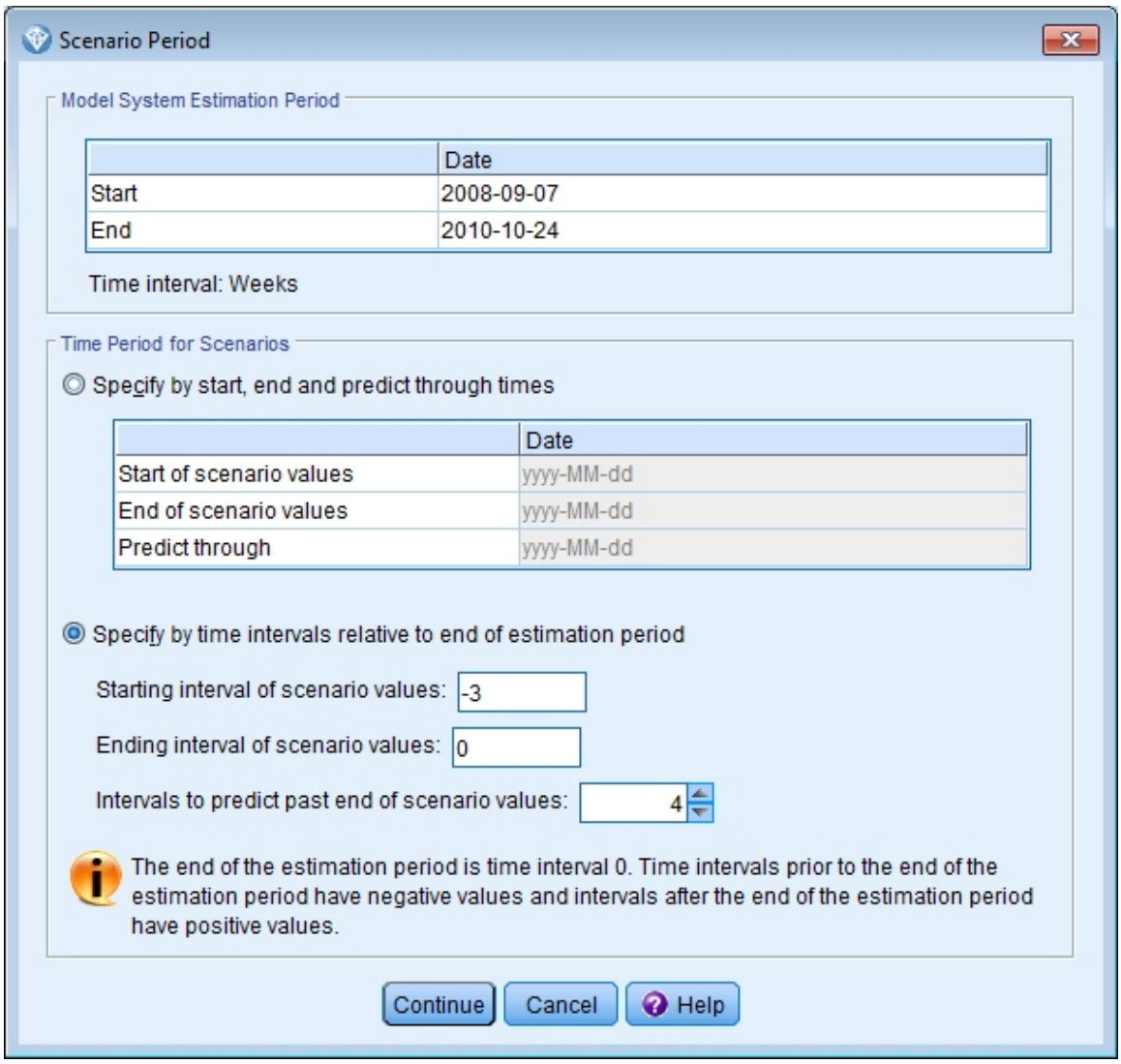


Figura 409. Período do Cenário

3. Selecione **Especificar por intervalos de tempo relativos ao fim do período de estimação**.
4. Insira `-3` para o intervalo inicial e insira `0` para o intervalo final.

   Essas configurações especificam que cada cenário é baseado em valores que são especificados para os últimos quatro intervalos de tempo no período de estimação. Por este exemplo, os últimos quatro intervalos de tempo significam as últimas quatro semanas. O intervalo de tempo sobre o qual os valores do cenário são especificados é referido como o *período de cenário*.
5. Insira 4 para os intervalos para prever após o término dos valores de cenários

   Essa configuração especifica que as previsões são geradas para quatro intervalos de tempo além do término do período do cenário.
6. Clique em **Continuar**.
7. Clique Em **Adicionar Cenário** na guia Cenários.

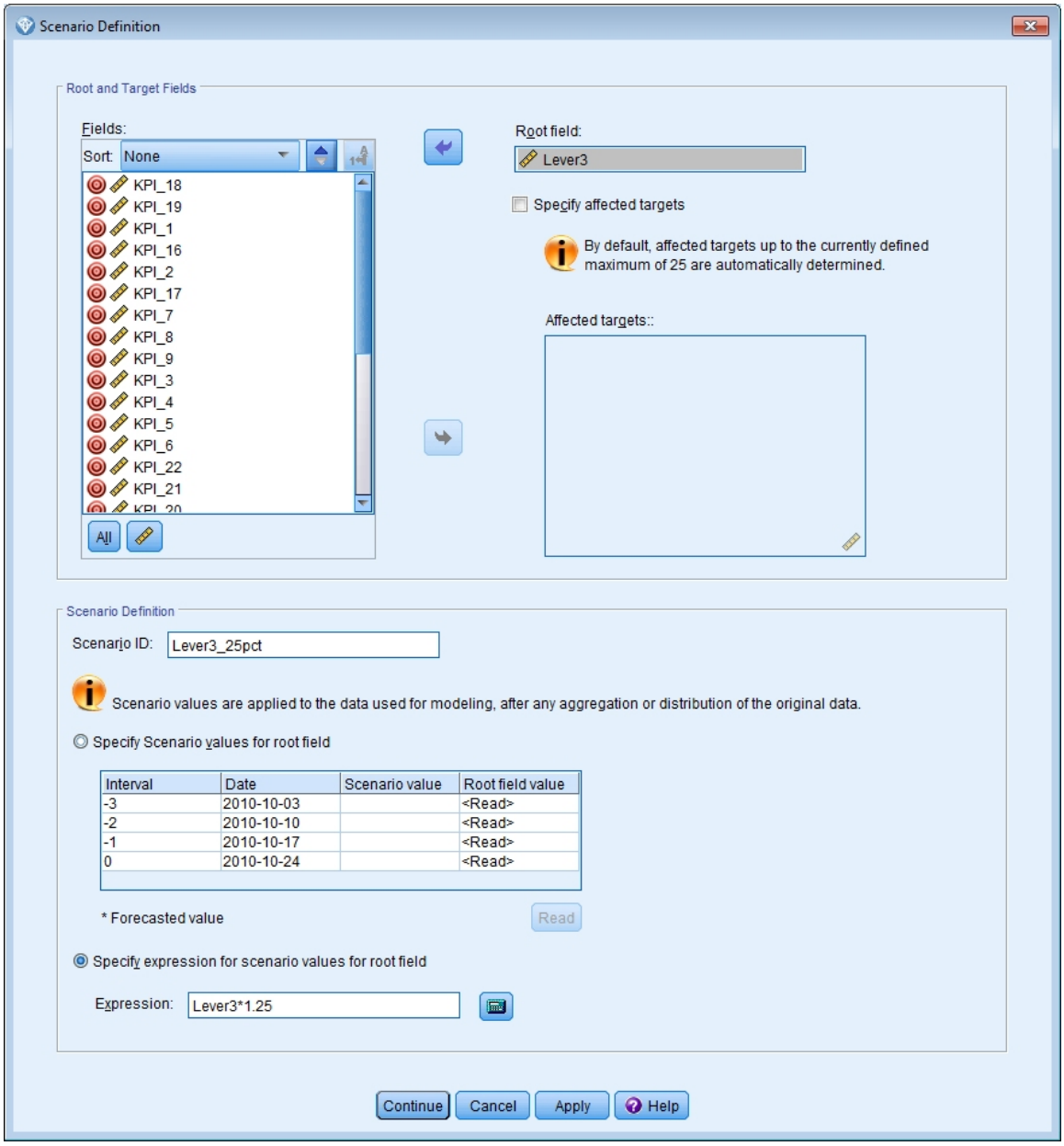


Figura 410. Definição de Cenário

8. Mova *Lever3* para a caixa **Campo Raiz** para examinar como valores especificados de *Lever3* no período do cenário afetam as previsões das outras séries que são afetadas causalmente por *Lever3*.
9. Insira `Lever3_25pct` para o ID do cenário.
10. Selecione **Especificar expressão para valores de cenário para campo raiz** e insira `Lever3*1.25` para a expressão.

    Esta configuração especifica que os valores para *Lever3* no período do cenário são 25% maiores do que os valores observados. Para expressões mais complexas, é possível utilizar o Expression Builder clicando no ícone da calculadora.
11. Clique em **Continuar**.
12. Repita as etapas 10 a 14 para definir um cenário que tenha *Lever3* para o campo raiz, `Lever3_50pct` para o ID do cenário e `Lever3*1.5` para a expressão.

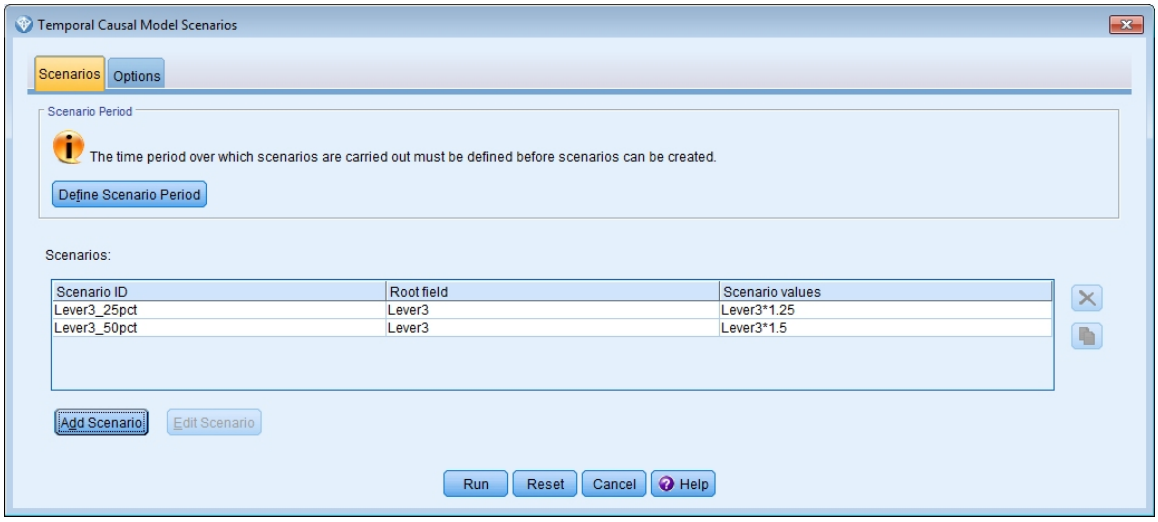
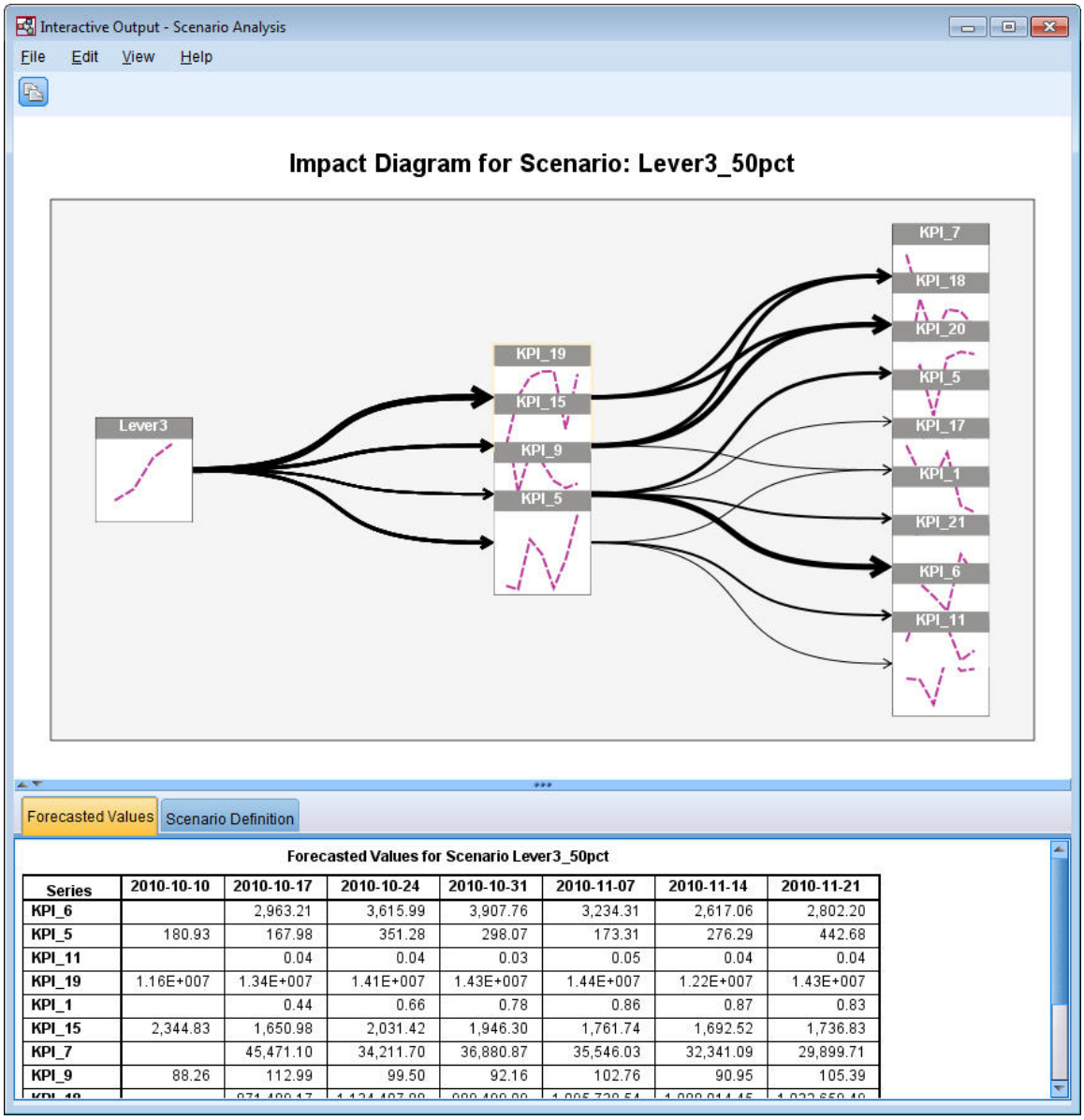


Figura 411. Cenários

13. Clique na guia **Opções** e digite *2* para o nível máximo para destinos afetados.
14. Clique em **Executar** .
15. Clique duas vezes no gráfico Diagrama de Impacto para *Lever3_50pct* no Viewer para ativá-lo.



Series	2010-10-10	2010-10-17	2010-10-24	2010-10-31	2010-11-07	2010-11-14	2010-11-21
KPI_6		2,963.21	3,615.99	3,907.76	3,234.31	2,617.06	2,802.20
KPI_5	180.93	167.98	351.28	298.07	173.31	276.29	442.68
KPI_11		0.04	0.04	0.03	0.05	0.04	0.04
KPI_19	1.16E+007	1.34E+007	1.41E+007	1.43E+007	1.44E+007	1.22E+007	1.43E+007
KPI_1		0.44	0.66	0.78	0.86	0.87	0.83
KPI_15	2,344.83	1,650.98	2,031.42	1,946.30	1,761.74	1,692.52	1,736.83
KPI_7		45,471.10	34,211.70	36,880.87	35,546.03	32,341.09	29,899.71
KPI_9	88.26	112.99	99.50	92.16	102.76	90.95	105.39

Figura 412. Diagrama De Impacto para Cenário: Lever3_50pct

O Diagrama de Impacto mostra a série que são afetadas pela série raiz *Lever3*. Dois níveis de efeitos são mostrados porque você especificou *2* para o nível máximo para destinos afetados.

A Tabela Valores Previstos inclui as previsões para todas as séries que são afetadas pelo *Lever3*, até o segundo nível de efeitos. As previsões para a série alvo no primeiro nível de efeitos começam no primeiro período de tempo após o início do período do cenário. Neste exemplo, as previsões para a série alvo no primeiro nível começam em 2010-10-10. As previsões para a série alvo no segundo nível de efeitos começam no segundo período de tempo após o início do período do cenário. Neste exemplo, as previsões para a série alvo no segundo nível começam em 2010-10-17. A natureza escalonada das previsões reflete o fato de que os modelos de séries temporais se baseiam em valores defesados dos insumos.

16. Clique no nó para *KPI_5* para gerar um diagrama de sequência detalhado.

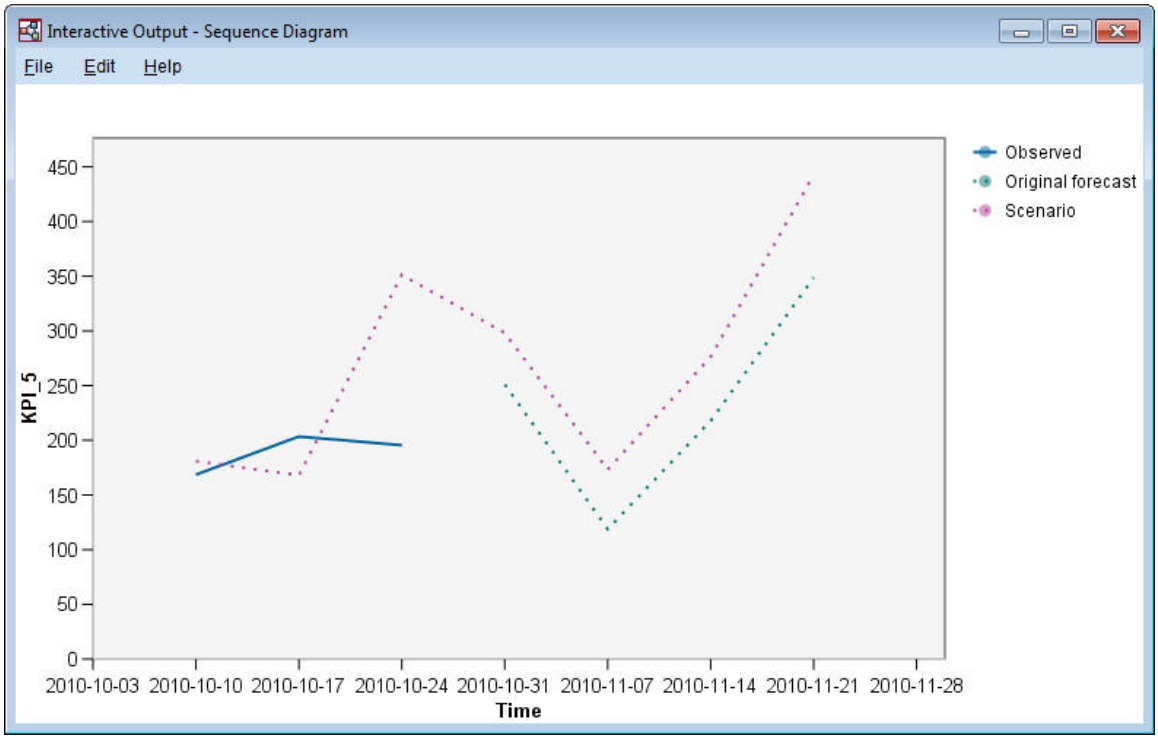


Figura 413. Diagrama Sequência para KPI_5

O gráfico de sequência mostra os valores previstos a partir do cenário, e também mostra os valores da série na ausência do cenário. Quando o período de cenário contém tempos dentro do período de estimação, os valores observados da série são mostrados. Por vezes além do fim do período de estimação, as previsões originais são mostradas.

Avisos

Estas informações foram desenvolvidas para produtos e serviços oferecidos nos EUA. Este material pode estar disponível na IBM em outros idiomas. Entretanto, poderá ser necessário ter uma cópia do produto ou uma versão do produto nesse idioma para acessá-lo.

É possível que a IBM não ofereça os produtos, serviços ou recursos discutidos nesta publicação em outros países. Consulte um representante IBM local para obter informações sobre produtos e serviços disponíveis atualmente em sua área. Qualquer referência a produtos, programas ou serviços IBM não significa que apenas produtos, programas ou serviços IBM possam ser utilizados. Qualquer outro produto, programa ou serviço, funcionalmente equivalente, poderá ser utilizado em substituição daqueles, desde que não infrinja nenhum direito de propriedade intelectual da IBM. Entretanto, a avaliação e verificação da operação de qualquer produto, programa ou serviço não IBM são de responsabilidade do Cliente.

A IBM pode ter patentes ou solicitações de patentes pendentes relativas a assuntos tratados nesta publicação. O fornecimento desta publicação não lhe garante direito algum sobre tais patentes. Pedidos de licença devem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil
IBM Brasil Ltda
Botafogo
Rio de Janeiro, RJ
CEP 22290-240

Para pedidos de licença relacionados a informações de DBCS (Conjunto de Caracteres de Byte Duplo), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie pedidos de licença, por escrito, para:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO “NO ESTADO EM QUE SE ENCONTRA”, SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS A ELAS NÃO SE LIMITANDO, AS GARANTIAS IMPLÍCITAS (OU CONDIÇÕES) DE NÃO INFRAÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias expressas ou implícitas em certas transações; portanto, essa disposição pode não se aplicar ao Cliente.

Essas informações podem conter imprecisões técnicas ou erros tipográficos. São feitas alterações periódicas nas informações aqui contidas; tais alterações serão incorporadas em futuras edições desta publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar os produtos e/ou programas descritos nesta publicação, sem aviso prévio.

Quaisquer referências nessas informações em sites não IBM são fornecidas por conveniência apenas e não de modo a servir como endosso para esses websites. Os materiais contidos nesses websites não fazem parte dos materiais desse produto IBM e a utilização desses websites é de inteira responsabilidade do Cliente.

A IBM pode usar ou distribuir quaisquer informações que você fornecer da forma como julgar apropriado, sem incorrer qualquer obrigação ao cliente.

Licenciados deste programa que desejam obter informações sobre este assunto com objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) a utilização mútua das informações trocadas, devem entrar em contato com:

Gerência de Relações Comerciais e Industriais da IBM Brasil
IBM Brasil Ltda
Botafogo
Rio de Janeiro, RJ
CEP 22290-240

Tais informações podem estar disponíveis, sujeitas a termos e condições apropriadas, incluindo em alguns casos o pagamento de uma taxa.

O programa licenciado descrito nesta publicação e todo o material licenciado disponível são fornecidos pela IBM sob os termos do Contrato com o Cliente IBM, do Contrato Internacional de Licença do Programa IBM ou de qualquer outro contrato equivalente.

Os dados de desempenho e exemplos de clientes citados são apresentados para propósitos ilustrativos somente. Os resultados do desempenho real podem variar, dependendo das configurações específicas e condições operacionais.

As informações a respeito de produtos não IBM foram obtidas de fornecedores dos próprios produtos, seus anúncios publicados ou outras fontes disponíveis publicamente. A IBM não testou esses produtos e não pode confirmar a precisão de seu desempenho ou de sua compatibilidade, nem quaisquer reclamações relacionadas aos produtos não IBM. Perguntas sobre os recursos de produtos não IBM devem ser endereçadas aos fornecedores desses produtos.

Instruções a respeito de direção ou intento futuro da IBM estão sujeitas a alteração ou retirada sem aviso, e representam somente metas ou objetivos.

Estas informações contêm exemplos de dados e relatórios utilizados nas operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos incluem nomes de indivíduos, empresas, marcas e produtos. Todos esses nomes são fictícios e quaisquer semelhanças a pessoas reais ou empresas é meramente coincidência.

Marcas comerciais

IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em várias jurisdições no mundo todo. Outros nomes de produto e de serviço podem ser marcas registradas da IBM ou de outras empresas. Uma lista atual de marcas comerciais da IBM está disponível na web em "Copyright and trademark information" em www.ibm.com/legal/copytrade.shtml.

Adobe, o logotipo Adobe, PostScript e o logotipo PostScript são marcas registradas ou marcas comerciais da Adobe Systems Incorporated nos Estados Unidos e/ou em outros países.

Intel, o logotipo Intel, Intel Inside, o logotipo Intel Inside, Intel Centrino, o logotipo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou de suas subsidiárias nos Estados Unidos e em outros países.

Linux é uma marca registrada da Linus Torvalds nos Estados Unidos, e/ou em outros países.

Microsoft, Windows, Windows NT e o logotipo Windows são marcas comerciais da Microsoft Corporation nos Estados Unidos e/ou em outros países.

UNIX é uma marca registrada da The Open Group nos Estados Unidos e em outros países.

Java e todas as marcas registradas e logotipos baseados em Java são marcas ou marcas registradas da Oracle e/ou suas afiliadas.

Termos e condições para documentação do produto

Permissões para uso dessas publicações são concedidas mas estão sujeitas aos termos e condições a seguir.

Aplicabilidade

Esses termos e condições estão completando quaisquer termos para uso do website IBM.

Uso pessoal

O Cliente pode reproduzir estas publicações para uso pessoal e não comercial, contanto que todos os avisos do proprietário sejam preservados. Você pode não distribuir, exibir ou fazer trabalhos derivados dessas publicações, ou de qualquer parte delas, sem o consentimento expresso da IBM.

Uso comercial

É possível reproduzir, distribuir e exibir estas publicações exclusivamente dentro de sua empresa desde que todos os avisos do proprietário sejam preservados. Você pode não fazer trabalhos derivados dessas publicações, ou reproduzir, distribuir ou exibir estas publicações ou qualquer parte delas fora de sua empresa, sem o consentimento expresso da IBM.

Direitos

Exceto quando expressamente concedida nessa permissão, nenhuma outra propriedade intelectual, licenças ou direitos são concedidos, nem expressos ou implícitos, às publicações ou quaisquer informações, dados, software ou outra propriedade intelectual contida nela.

A IBM reserva-se ao direito de retirar as permissões concedidas aqui sempre que, de acordo com seus critérios, o uso das publicações for prejudicial ao seu interesse ou, conforme determinado pela IBM, as instruções acima não estiverem sendo seguidas adequadamente.

Você não pode fazer download, exportar ou re-exportar essas informações, exceto em conformidade total com todas as leis e regulamentações aplicáveis, incluindo todas as leis e regulamentações de exportação dos Estados Unidos.

IBM NÃO APRESENTA QUALQUER GARANTIA SOBRE O CONTEÚDO DESSAS PUBLICAÇÕES. AS PUBLICAÇÕES SÃO FORNECIDAS "COMO ESTÃO" E SEM GARANTIA DE QUALQUER TIPO, EXPRESSAS OU IMPLÍCITAS, INCLUINDO MAS NÃO SE LIMITANDO A GARANTIAS IMPLÍCITAS DE COMERCIALIZAÇÃO, NÃO INFRAÇÃO E ADEQUAÇÃO A UM DETERMINADO PROPÓSITO.

Index

G

I

J

L

M

N

P

Q

R

S

T

U

V