Guia do IBM SPSS Modeler CRISP-DM

Nota

Antes de utilizar essas informações e o produto que elas suportam, leia as informações em “Avisos” na página 39.

Informações do produto

Esta edição se aplica à versão 18, release 4, modificação 0 de IBM® SPSS Modeler e a todos os lançamentos e modificações subsequentes até indicado de outra forma em novas edições.

© Copyright International Business Machines Corporation .

Índice

Prefácio

IBM SPSS Modeler é o ambiente de trabalho de mineração de dados de força corporativa do IBM Corp. . O SPSS Modeler ajuda as organizações a melhorarem as relações com o cliente e com o cidadão por meio de um entendimento profundo dos dados. As organizações utilizam o insight adquirido do SPSS Modeler para reter clientes rentáveis, identificar oportunidades de venda cruzada, atrair novos clientes, detectar fraude, reduzir o risco e melhorar a entrega de serviço de governo.

A interface visual do SPSS Modeler convida os usuários a aplicarem seus conhecimentos de negócios específicos, levando a modelos preditivos mais poderosos e reduzindo o tempo para a solução. O SPSS Modeler oferece muitas técnicas de modelagem, como predição, classificação, segmentação e algoritmos de detecção de associação. Quando os modelos são criados, o IBM SPSS Modeler Solution Publisher permite entregá-los aos tomadores de decisão na empresa ou a um banco de dados.

Sobre o IBM Business Analytics

O software IBM Business Analytics fornece informações completas, consistentes e exatas nas quais os tomadores de decisão confiam para melhorar o desempenho de negócios. Um portfólio abrangente de inteligência de negócios, analise preditiva, gerenciamento de desempenho financeiro e estratégiae aplicativos analíticos fornecem insight claro, imediato e prático sobre o desempenho atual e a capacidade de prever resultados futuros. Combinado com soluções para segmentos do mercado, práticas comprovadas e serviços profissionais completos, organizações de qualquer tamanho poderão conduzir maior produtividade, automatizar as decisões de modo confiável e entregar melhores resultados.

Como parte deste dossier, o software IBM SPSS Predictive Analytics ajuda as organizações a prever futuros eventos e agir proativamente com esse insight para melhores resultados de negócios. Os clientes acadêmicos, comerciais e do governo no mundo todo se baseiam na tecnologia do IBM SPSS como uma vantagem competitiva para atrair, manter e aumentar seus clientes, enquanto reduz fraudes e minimiza riscos. Ao incorporar o software IBM SPSS em suas operações diárias, as organizações tornam-se empreendimentos preditivos-capazes de direcionar e automatizar as decisões para cumprir metas de negócios e obter vantagem competitiva mensurável. Para obter informações adicionais ou para entrar em contato com um representante, visite http://www.ibm.com/spss.

Suporte técnico

O suporte técnico está disponível para manutenção dos clientes. Os clientes podem entrar em contato com o Suporte Técnico para assistência no uso de produtos IBM ou para ajuda de instalação para um dos ambientes de hardware suportados. Para chegar ao Suporte Técnico, consulte o site da IBM em http://www.ibm.com/support. Esteja preparado para se identificar, sua organização e seu contrato de suporte ao solicitar assistência.

Capítulo 1. Introdução ao CRISP-DM

Visão geral da ajuda do CRISP-DM

CRISP-DM, que significa Processo Padrão de Vários Segmentos de Mercados para Mineração de Dados, é uma forma comprovada pelo mercado para orientar seus esforços de mineração de dados.|

- Como uma **metodologia**, ela inclui descrições das fases típicas de um projeto, as tarefas envolvidas em cada fase e uma explicação dos relacionamentos entre essas tarefas.
- Como um **modelo de processo**, o CRISP-DM fornece uma visão geral do ciclo de vida da mineração de dados.

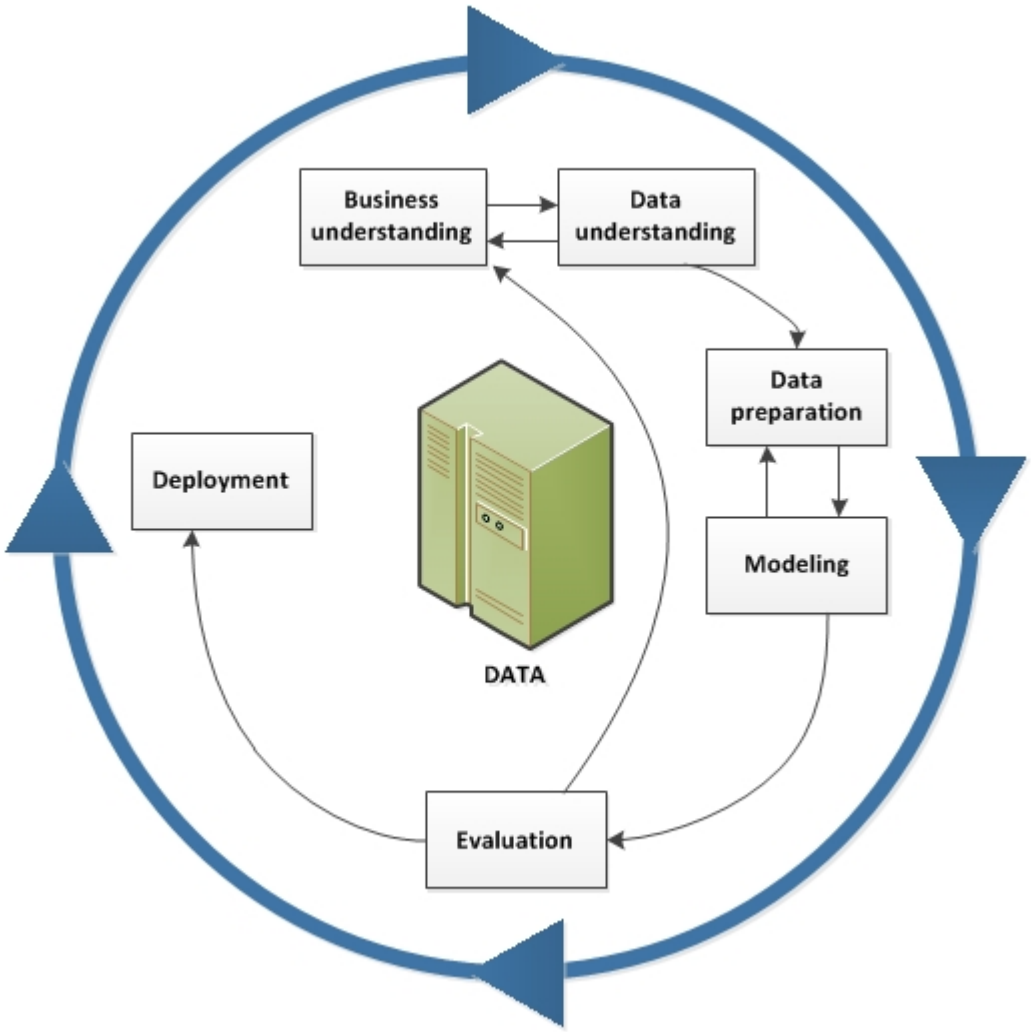

Figura 1. O ciclo de vida da mineração de dados

O modelo de ciclo de vida é composto de seis fases com setas indicando as dependências mais importantes e frequentes entre as fases. A sequência das fases não é rigorosa. De fato, a maioria dos projetos vão e voltam entre as fases, conforme necessário.

O modelo CRISP-DM é flexível e pode ser facilmente customizado. Por exemplo, se sua organização planejar detectar a lavagem de dinheiro, é provável que você examine detalhadamente grandes quantidades de dados sem uma meta de modelagem específica. Em vez da modelagem, seu trabalho irá se concentrar na exploração e visualização de dados para descobrir os padrões suspeitos em dados

financeiros. O CRISP-DM permite que você crie um modelo de mineração de dados que se encaixe em suas necessidades específicas.

Em tal situação, as fases de modelagem, avaliação e implementação podem ser menos relevantes do que as fases de entendimento e preparação de dados. Entretanto, ainda é importante considerar algumas das questões levantadas durante essas fases posteriores para o planejamento de longo prazo e para futuras metas de mineração de dados.

CRISP-DM em IBM SPSS Modeler

O IBM SPSS Modeler incorpora a metodologia do CRISP-DM em duas formas para fornecer um suporte exclusivo para a mineração de dados efetiva.

- A ferramenta do projeto do CRISP-DM ajuda a organizar os fluxos, a saída e as anotações do projeto, de acordo com as fases de um típico projeto de mineração de dados. É possível produzir os relatórios a qualquer momento durante o projeto com base nas notas para fluxos e fases do CRISP-DM.
- A ajuda do CRISP-DM o orienta no processo de condução de um projeto de mineração de dados. O sistema de ajuda inclui listas de tarefas para cada etapa, bem como exemplos de como o CRISP-DM funciona no mundo real. É possível acessar a Ajuda do CRISP-DM escolhendo **Ajuda do CRISP-DM** no menu Ajuda da janela principal.

Ferramenta do projeto do CRISP-DM

A ferramenta do projeto do CRISP-DM fornece uma abordagem estrutura para a mineração de dados que podem ajudar a garantir o sucesso de seu projeto. Ela é, essencialmente, uma extensão da ferramenta de projeto do IBM SPSS Modeler. De fato, é possível alternar entre a visualização do CRISP-DM e a visualização de Classes padrão para ver os seus fluxos e a saída organizada por tipo ou por fases do CRISP-DM.

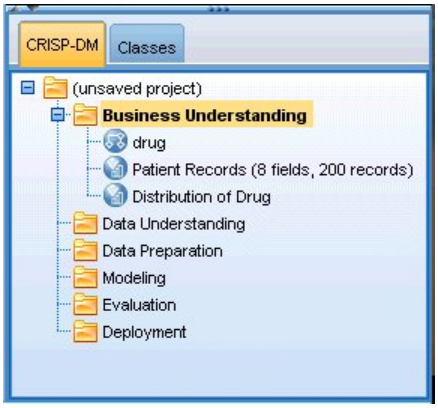

Figura 2. Ferramenta do projeto do CRISP-DM

Ao usar a visualização do CRISP-DM da ferramenta do projeto, é possível:

- Organize os fluxos e a saída de um projeto de acordo com as fases de mineração de dados.
- Tome notas das metas de sua organização para cada fase.
- Crie dicas de ferramenta customizadas para cada fase.
- Tome notas sobre as conclusões tiradas de um gráfico ou modelo específico.
- Gere um relatório em HTML ou atualize para distribuição para a equipe do projeto.

Ajuda do CRISP-DM

O IBM SPSS Modeler oferece um guia on-line para o modelo de processo não proprietário do CRISP-DM. O guia é organizado pelas fases do projeto e fornece o seguinte suporte:

- Uma visão geral e uma lista de tarefas para cada fase do CRISP-DM

- Ajuda na produção de relatórios para diversos marcos
- Exemplos do mundo real que ilustram como uma equipe de projeto pode usar o CRISP-DM para facilitar o caminho para a mineração de dados
- Links para recursos adicionais sobre o CRISP-DM

É possível acessar a Ajuda do CRISP-DM escolhendo **Ajuda do CRISP-DM** no menu Ajuda da janela principal.

Recursos adicionais

Além do suporte do IBM SPSS Modeler para o CRISP-DM, há diversas formas de expandir seu entendimento de processos de mineração de dados.

- Leia o manual do CRISP-DM, criado pelo consórcio do CRISP-DM e fornecido com esta liberação.
- Read *Data Mining with Confidence*, copyright 2002 por SPSS Inc., ISBN 1-56827-287-1.

Capítulo 2. Entendimento de Negócios

Visão geral do entendimento dos negócios

Mesmo antes de trabalhar no IBM SPSS Modeler, você deve tirar algum tempo para explorar aquilo que sua organização espera obter da mineração de dados. Tente envolver tantas pessoas-chave quanto possível nessas discussões e documente os resultados. A etapa final dessa fase do CRISP-DM discute como produzir um plano de projeto usando as informações aqui reunidas.

Embora esta pesquisa pareça indispensável, ela não é. Começar a conhecer os motivos dos negócios para o seu esforço na mineração de dados ajuda a garantir que todos estejam na mesma página antes de gastar recursos valiosos.

Determinando os objetivos de negócios

A sua primeira tarefa é tentar obter tanta percepção quanto possível das metas de negócios para a mineração de dados. Isso pode não ser tão fácil quanto parece, mas é possível minimizar riscos posteriores esclarecendo problemas, metas e recursos.

A metodologia CRISP-DM fornece uma forma estruturada para realizar isso.

Lista de tarefas

- Comece a reunir informações básicas sobre a situação atual de negócios.
- Objetivos de negócios específicos do documento decididos pelos principais tomadores de decisões.
- Concorde com os critérios usados para determinar o sucesso da mineração de dados a partir da perspectiva do negócio.

Exemplo de varejo eletrônico--Descobrindo os objetivos de negócios

Um cenário de mineração na web usando o CRISP-DM

À medida que mais empresas fazem a transição para as vendas na Web, um varejista eletrônico estabelecido de computadores/eletroeletrônicos enfrenta uma concorrência crescente de novos sites. Diante da realidade de que as lojas na Web estão surgindo tão rapidamente (ou mais rapidamente!) quanto os clientes estão migrando para a Web, a empresa deve encontrar maneiras de permanecer rentável, apesar dos custos crescentes da aquisição do cliente. Uma solução proposta é cultivar os relacionamentos existentes com o cliente para maximizar o valor de cada um dos clientes atuais da empresa.

Assim, um estudo foi encomendado com os seguintes objetivos:

- Melhorar as vendas cruzadas fazendo melhores recomendações.
- Aumentar a fidelidade do cliente com um serviço mais personalizado.

Como tentativa, o estudo será julgado um sucesso se:

- As vendas cruzadas aumentarem em 10%.
- Os clientes passarem mais tempo e consultarem mais páginas no site por visita.
- O estudo for concluído em tempo e abaixo do orçamento.

Compilando o segundo plano dos negócios

Entender a situação de negócios de sua organização o ajuda a saber com o que você está trabalhando em termos de:

- Recursos disponíveis (equipe e material)

- Problemas
- Objetivos

Será necessário pesquisar um pouco sobre a situação de negócios atual, a fim de encontrar respostas reais para perguntas que podem impactar no resultado do projeto de mineração de dados.

Tarefa 1--Determinar estrutura organizacional

- Desenvolver gráficos organizacionais para ilustrar as divisões, os departamentos e os grupos de projetos corporativos. Certifique-se de incluir nomes e responsabilidades.
- Identificar os principais indivíduos na organização.
- Identificar um patrocinador interno que fornecerá o suporte e/ou o conhecimento de domínio.
- Determinar se há um comitê diretor e obter uma lista de membros.
- Identificar unidades de negócios que serão afetadas pelo projeto de mineração de dados.

Tarefa 2--Descrever a área do problema

- Identificar a área do problema, como marketing, atendimento ao cliente ou desenvolvimento de negócios.
- Descrever o problema em termos gerais.
- Esclarecer os pré-requisitos do projeto. Quais são as motivações por trás do projeto? Os negócios já usam a mineração de dados?
- Verificar o status do projeto de mineração de dados no grupo de negócios. O esforço já foi aprovado ou a mineração de dados precisa ser "divulgada" como a tecnologia principal para o grupo de negócios?
- Se necessário, prepare apresentações informativas sobre a mineração de dados para sua organização.

Tarefa 3--Descrever a solução atual

- Descrever qualquer solução usada atualmente para abordar os problemas de negócios.
- Descrever as vantagens e desvantagens da solução atual. Além disso, abordar o nível de aceitação que esta solução teve na organização.

Definindo objetivos de negócios

Aqui é onde as coisas passam a ser específicas. Como resultado de sua pesquisa e de suas reuniões, você deve construir um objetivo primário concreto conforme combinado pelos patrocinadores do projeto e outras unidades de negócios afetadas pelos resultados. Essa meta eventualmente será convertida de algo tão nebuloso quanto a "redução da perda de clientes" para os objetivos específicos de mineração de dados que orientará suas analíticas.

Lista de tarefas

Certifique-se de tomar notas sobre os seguintes pontos para posterior incorporação no plano do projeto. Lembre-se de manter as metas realísticas.

- Descreva o problema que deseja solucionar usando a mineração de dados.
- Especifique todas as questões de negócios tão precisamente quanto possível.
- Determine qualquer outra necessidade de negócios (como não perder nenhum cliente existente enquanto aumenta as oportunidades de venda cruzada).
- Especifique os benefícios esperados em termos de negócios (tal como a redução da perda de clientes entre clientes de alto valor em 10%).

Critérios do sucesso dos negócios

A meta à frente pode estar clara, mas você saberá assim que chegar lá? É importante definir a natureza do sucesso dos negócios para seu projeto de mineração de dados antes de continuar. Os critérios de sucesso incidem em duas categorias:

- **Objetivos.** Esses critérios podem ser tão simples quando um aumento específico na precisão de auditorias ou uma redução concordada na perda de clientes.
- **Subjetivos.** Os critérios subjetivos como "descobrir grupos de tratamentos eficazes" são mais difíceis de serem atendidos, mas é possível combinar quem fará a decisão final.

Lista de tarefas

- Documente, tão precisamente quanto possível, os critérios de sucesso para esse projeto.
- Certifique-se de cada objeto de negócios ter um critério correlativo para sucesso.
- Alinhe os árbitros das medições subjetivas do sucesso. Se possível, tome notas de suas expectativas.

Avaliando a situação

Agora que você tem uma meta claramente definida, é hora de fazer uma avaliação de onde você está neste momento. Esta etapa envolve fazer perguntas como:

- Quais tipos de dados estão disponíveis para análise?
- Você tem a equipe necessária para concluir o projeto?
- Quais são os maiores fatores de risco envolvidos?
- Você tem um plano de contingência para cada risco?

Exemplo de varejo eletrônico--Avaliando a situação

Um cenário de mineração na web usando o CRISP-DM

Esta é a primeira tentativa do varejista eletrônico de eletroeletrônico em mineração na Web e a empresa decidiu consultar um especialista em mineração de dados para ajudar na introdução. Uma das primeiras tarefas com a qual o consultor se depara é a avaliação dos recursos da empresa para a mineração de dados.

Equipe. Está claro que há conhecimento interno em relação a gerenciamento de logs do servidor e de bancos de dados de produtos e compras, mas pouca experiência em data warehouse e na limpeza de dados para análise. Assim, um especialista em banco de dados também deve ser consultado. Visto que a empresa espera que os resultados do estudo se tornem uma parte de um processo contínuo de mineração na web, o gerenciamento também deve levar em conta se qualquer posição criada durante o esforço atual irá se tornar permanente.

Dados. Visto que esta é uma empresa estabelecida, há logs da Web e dados de compras suficientes nos quais se basear. De fato, para este estudo inicial, a empresa restringirá a análise a clientes que se "registraram" no site. Se for bem-sucedido, o programa poderá ser expandido.

Riscos. Além dos dispêndios monetários para os consultores e o tempo gasto pelos funcionários no estudo, não há um grande risco imediado neste empreendimento. Entretanto, o tempo é sempre importante, então este projeto inicial é planejado para um único trimestre financeiro.

Além disso, não há um grande fluxo de caixa extra no momento, portanto é indispensável que o estudo ocorra abaixo do orçamento. Caso uma dessas metas esteja em risco, os gerentes de negócios sugerem que o escopo do projeto seja reduzido.

Inventário do recurso

Fazer um inventário preciso de seus recursos é indispensável. É possível economizar muito tempo e dores de cabeça observando atentamente o hardware, as fontes de dados e os problemas com a equipe.

Tarefa 1--Pesquisar recursos de hardware

- De qual hardware precisa para o suporte?

Tarefa 2--Identificar as fontes de dados e os depósitos de conhecimento

- Quais fontes de dados estão disponíveis para a mineração de dados? Tome nota dos tipos e formatos de dados.

- Como os dados são armazenados? Você tem acesso em tempo real aos data warehouses ou aos bancos de dados operacionais?
- Você planeja adquirir mais dados externos, como informações demográficas?
- Há problemas de segurança que impedem o acesso aos dados necessários?

Tarefa 3--Identificar recursos de equipe

- Você tem acesso a especialistas de dados e de negócios?
- Você identificou administradores de base de dados e outra equipe de suporte que possam ser necessários?

Uma vez que tenha feito essas perguntas, inclua uma lista de contatos e recursos no relatório da fase.

Requisitos, suposições e restrições

Os seus esforços terão mais chances de serem compensados se fizer uma avaliação honesta das responsabilidades do projeto. Tornar essas preocupações tão explícitas quanto possível ajudará a evitar problemas futuros.

Tarefa 1--Determinar os requisitos

O requisito fundamental é a meta de negócio discutida anteriormente, mas considere o seguinte:

- Há restrições de segurança e legais em relação aos dados ou aos resultados do projeto?
- Todos estão alinhados com os requisitos de planejamento do projeto?
- Há requisitos para a implementação de resultados (por exemplo, publicar na Web ou ler os escores em um banco de dados)?

Tarefa 2--Esclarecer suposições

- Existem fatores econômicos que podem afetar o projeto (por exemplo, consultar taxas ou produtos competitivos)?
- Existem suposições sobre a qualidade de dados?
- Como a equipe de patrocinador/gerenciamento do projeto esperam visualizar os resultados? Em outras palavras, eles desejam compreender o próprio modelo ou simplesmente visualizar os resultados?

Tarefa 3--Verificar restrições

- Você tem todas as senhas necessária para o acesso aos dados?
- Você verificou todas as restrições legais sobre o uso dos dados?
- Todas as restrições financeiras estão abrangidas no orçamento do projeto?

Riscos e contingências

Também é sensato considerar os possíveis riscos no decorrer do projeto. Os tipos de risco incluem:

- Planejamento (E se o projeto levar mais tempo do que o previsto?)
- Financeiro (E se o patrocinador do projeto se deparar com problemas orçamentários?)
- Dados (E se os dados forem de qualidade ou cobertura insatisfatória?)
- Resultados (E se os resultados iniciais forem menos dramáticos do que o esperado?)

Depois de considerar os diversos riscos, forneça um plano de contingência para ajudar a evitar um desastre.

Lista de tarefas

- Documente cada risco possível.
- Documente um plano de contingência para cada risco.

Terminologia

Para garantir que as equipes de negócios e de mineração de dados estejam "falando o mesmo idioma", é necessário considerar a compilação de um glossário de termos técnicos e de palavras de efeito que precisam de esclarecimento. Por exemplo, se "perda de clientes" para sua empresa tem um significado específico e exclusivo, vale a pena indicar isso explicitamente para o proveito de toda a equipe. Da mesma forma, a equipe pode se beneficiar do esclarecimento do uso de um gráfico de ganhos.

Lista de tarefas

- Mantenha uma lista de termos ou jargões confusos para os membros da equipe. Inclua a terminologia de negócios e de mineração de dados.
- Considere publicar a lista na intranet ou na documentação do projeto.

Análise de custo-benefício

Esta etapa responde a pergunta: **Qual é o seu resultado?** Como parte da avaliação final, é essencial comparar os custos do projeto aos potenciais benefícios do sucesso.

Lista de tarefas

Inclua em sua análise os custos estimados para:

- Coleta de dados e quaisquer dados externos usados
- Implementação de resultados
- Custos operacionais

Em seguida, leve em consideração os benefícios de:

- Cumprimento do objetivo primário
- Percepções adicionais geradas a partir da exploração de dados
- Possíveis benefícios da melhor compreensão de dados

Determinando objetivos para mineração de dados

Agora que a meta de negócio está clara, é hora de convertê-la em uma realidade de mineração de dados. Por exemplo, o objetivo de negócios para "perda de clientes" pode ser convertido em um objetivo de mineração de dados que inclui:

- Identificar clientes de alto valor com base nos dados recentes de compra
- Construir um modelo usando dados disponíveis do cliente para prever a probabilidade de migração para o concorrente de cada cliente
- Designar a cada cliente um posto baseado na propensão de perda de clientes e de valor do cliente

Esses objetivos de mineração de dados, se atendidos, podem ser usados pelos negócios para reduzir a perda de clientes entre os clientes mais valiosos.

Como se pode ver, negócios e tecnologia trabalham lado a lado para uma mineração de dados efetiva. Continue a ler para obter dicas específicas sobre como determinar os objetivos de mineração de dados.

Metas de mineração de dados

À medida que você trabalha com analistas de negócios e de dados para definir uma solução técnica para os problemas de negócios, lembre-se de manter as coisas concretas.

Lista de tarefas

- Descreva o **tipo** de problema da mineração de dados, como armazenamento em cluster, predição ou classificação.
- Documente as metas técnicas usando unidades específicas de tempo, como predições com uma validade de três meses.

- Se possível, forneça números reais para os resultados desejados, como a produção de escores de perda de clientes para 80% dos clientes existentes.

Exemplo de varejo eletrônico--Metas da mineração de dados

Um cenário de mineração na web usando o CRISP-DM

Com a ajuda de seu consultor de mineração de dados, o varejista eletrônico foi capaz de converter os objetos de negócios da empresa em termos de mineração de dados. As metas para que o estudo inicial seja concluído neste trimestre são:

- Use informações históricas sobre compras anteriores para gerar um modelo que vincule itens "relacionados". Quando os usuários olham uma descrição de item, forneça vínculos para outros itens no grupo relacionado (**análise de cesta de compras**).
- Use logs da Web para determinar aquilo que diferentes clientes estão tentando encontrar e, então, projete novamente o site para destacar esses itens. Cada "tipo" de cliente diferente verá uma página principal diferente para o site (**determinação de perfil**).
- Use logs da Web para tentar prever para onde a pessoa irá a seguir, dado de onde ela veio e se já esteve em seu site (**análise de sequência**).

Critérios de sucesso da mineração de dados

O sucesso também deve ser definido em termos de manter os esforços de sua mineração de dados sob controle. Use a meta de mineração de dados determinada anteriormente para formular avaliações de desempenho para obter sucesso. O IBM SPSS Modeler fornece ferramentas como o nó de Avaliação e o nó de Análise para ajudá-lo a analisar a precisão e a validade de seus resultados.

Lista de tarefas

- Descrever os métodos para avaliação do modelo (por exemplo, precisão, desempenho, etc.).
- Defina avaliações de desempenho para avaliar o sucesso. Forneça números específicos.
- Defina medidas subjetivas tão bem quanto possível e determine o árbitro do sucesso.
- Considere se a implementação bem-sucedida dos resultados do modelo é parte do sucesso da mineração de dados. Comece a planejar a implementação agora.

Produzindo um plano do projeto

Neste ponto, você está pronto para produzir um plano para o projeto de mineração de dados. As perguntas feitas até agora e as metas de negócios e mineração de dados formuladas formarão a base deste roteiro.

Escrevendo o plano do projeto

O plano do projeto é o documento principal para todo o seu trabalho de mineração de dados. Se bem feito, ele poderá informar todos aqueles associados ao projeto de metas, recursos, riscos e planejamento de todas as fase da mineração de dados. É possível que queira publicar o plano, bem como a documentação reunida em toda esta fase na intranet de sua empresa.

Lista de tarefas

Ao criar o plano, certifique-se de ter respondido as seguintes perguntas:

- Você discutiu as tarefas do projeto e o plano proposto com todos os envolvidos?
- As estimativas de tempo foram incluídas em todas as fases ou tarefas?
- Você incluiu o esforço ou os recursos necessários para implementar os resultados para a solução de negócios?
- Os pontos de decisão e as solicitações de revisão estão destacados no plano?
- Você marcou as fases em que normalmente ocorrem diversas iterações, como modelagem?

Plano do projeto de amostra

O plano de visão geral do estudo é como este mostrado na tabela abaixo.

Tabela 1. Visão geral do plano do projeto de amostra

Fase	Hora	Recursos	Riscos
Entendimento dos negócios	1 semana	Todos os analistas	Mudança econômica
Entendimento de dados	3 semanas	Todos os analistas	Problemas de dados, problemas de tecnologia
Preparação de dados	5 semanas	Consultor de mineração de dados, algum tempo como analista de banco de dados	Problemas de dados, problemas de tecnologia
Modelagem	2 semanas	Consultor de mineração de dados, algum tempo como analista de banco de dados	Problemas de tecnologia, incapacidade de localizar um modelo adequado
Avaliação	1 semana	Todos os analistas	Mudança econômica, incapacidade de implementar os resultados
Implementação	1 semana	Consultor de mineração de dados, algum tempo como analista de banco de dados	Mudança econômica, incapacidade de implementar os resultados

Avaliando ferramentas e técnicas

Visto que você já escolheu usar o IBM SPSS Modeler como sua ferramenta para o sucesso da mineração de dados, é possível usar esse passo para pesquisar quais técnicas de mineração de dados são mais apropriadas para suas necessidades de negócios. O IBM SPSS Modeler oferece uma gama completa de ferramentas para cada fase da mineração de dados. Para decidir quando usar as diversas técnicas, consulte a seção de modelagem da Ajuda on-line.

Pronto para o próximo passo?

Antes de explorar os dados e começar a trabalhar no IBM SPSS Modeler, certifique-se de ter respondido as seguintes perguntas.

De uma perspectiva do negócio:

- O que seu negócio espera alcançar com este projeto?
- Como você definirá a conclusão bem-sucedida de nossos esforços?
- Você tem o orçamento e os recursos necessários para atingir suas metas?
- Você tem acesso a todos os dados necessários para este projeto?
- Você e sua equipe já discutiram os riscos e as contingências associados a este projeto?
- Os resultados de sua análise de custo-benefício tornam este projeto vantajoso?

Depois de ter respondido as perguntas acima, você converteu essas respostas em uma meta de mineração de dados?

De uma perspectiva da mineração de dados:

- Especificamente, como a mineração de dados pode ajudá-lo a atingir suas metas de negócios?
- Você tem alguma ideia de qual técnica de mineração de dados pode produzir os melhores resultados?
- Como você saberá quando os resultados são precisos ou eficazes o suficiente? (*Foi definida uma medida do sucesso da mineração de dados?*)
- Como os resultados da modelagem foram implementados? Você levou em consideração a implementação em seu plano do projeto?
- O plano do projeto inclui todas as fases do CRISP-DM?
- Os riscos e as dependências foram considerados no plano?

Se você respondeu "sim" às perguntas acima, então você está pronto para olhar mais atentamente os dados.

Capítulo 3. Entendimento de Dados

Visão geral do entendimento de dados

A fase de entendimento de dados do CRISP-DM envolve olhar mais atentamente os dados disponíveis para mineração. Esse passo é essencial para evitar problemas inesperados durante a etapa seguinte, a preparação de dados, que é normalmente a parte mais longa de um projeto.

O entendimento de dados envolve acessar os dados e explorá-los usando tabelas e gráficos que podem ser organizados no IBM SPSS Modeler usando a ferramenta de projeto CRISP-DM. Isso permite determinar a qualidade dos dados e descrever os resultados dessas etapas na documentação do projeto.

Coletando dados iniciais

Neste ponto no CRISP-DM, você está pronto para acessar os dados e levá-los para o IBM SPSS Modeler. Os dados chegam de uma variedade de fontes, tais como:

- **Dados existentes.** Isso inclui uma grande variedade de dados, como dados transacionais, dados de pesquisa, logs da Web, etc. Considere se os dados existentes são suficientes para atender às suas necessidades.
- **Dados comprados.** Sua organização usa dados suplementares, como demográficos? Em caso negativo, leve em consideração se podem ser necessários.
- **Dados adicionais.** Se as fontes acima não atenderem suas necessidades, poderá ser necessário fazer pesquisas de opinião ou iniciar o rastreamento adicional para suplementar os armazenamentos de dados existentes.

Lista de tarefas

Observe os dados no IBM SPSS Modeler e considere as seguintes questões. Certifique-se de tomar notas de suas descobertas. Veja o tópico “Escrevendo um relatório de coleta de dados” na página 14 para obter mais informações.

- Quais atributos (colunas) do banco de dados parecem mais promissores?
- Quais atributos parecem irrelevantes e podem ser excluídos?
- Há dados suficientes para tirar conclusões generalizáveis ou fazer predições precisas?
- Há atributos em excesso para o seu método de modelagem escolhido?
- Você está mesclando diversas fontes de dados? Em caso positivo, há áreas que podem apresentar um problema ao mesclar?
- Você levou em consideração como os valores omissos são manipulados em cada uma de suas fontes de dados?

Exemplo de varejo eletrônico--Coleta inicial de dados

Um cenário de mineração na web usando o CRISP-DM

O varejista eletrônico neste exemplo usa diversas fontes de dados importantes, incluindo:

Logs da Web. Os logs de acesso brutos contêm todas as informações sobre como os clientes navegam no website. As referências a arquivos de imagem e a outras entradas não informativas nos logs da Web precisarão ser removidas como parte do processo de preparação de dados.

Dados de compra. Quando um cliente envia um pedido, todas as informações pertinentes a essa ordem são salvas. As ordens no banco de dados de compra precisam ser mapeados para as sessões correspondentes nos logs da Web.

Banco de dados do produto. Os atributos do produto podem ser úteis ao determinar os produtos "relacionados". As informações do produto precisam ser mapeadas para as ordens correspondentes.

Banco de dados do cliente. Este banco de dados contém informações adicionais coletadas de clientes registrados. Os registros não estão completos de forma alguma, pois vários clientes não preenchem os questionários. As informações do cliente precisam ser mapeadas para as compras e sessões correspondentes nos logs da Web.

Nesse momento, a empresa não tem nenhum plano de comprar bancos de dados externos ou de gastar dinheiro fazendo pesquisas, pois seus analistas estão ocupados gerenciando os dados que têm atualmente. Em algum momento, entretanto, eles podem desejar considerar uma implementação estendida de resultados da mineração de dados, caso no qual a compra de dados demográficos adicionais para clientes não registrados pode ser muito útil. Pode ser útil ter informações demográficas para ver como a base do cliente do varejista eletrônico se difere do comprador médio da Web.

Escrevendo um relatório de coleta de dados

Ao usar o material reunido no passo anterior, é possível começar a escrever o relatório de coleta de dados. Assim que estiver concluído, o relatório poderá ser incluído no site da Web ou distribuído para a equipe. Ele também pode ser combinado com os relatórios preparados nos passos seguintes: descrição, exploração e verificação de qualidade dos dados. Esses relatórios irão orientar seu trabalho durante toda a fase de preparação de dados].

Descrevendo dados

Há diversas formas de descrever os dados, mas a maioria das descrições se concentra na quantidade e na qualidade dos dados: quantos dados estão disponíveis e a condição dos dados. Listados abaixo estão algumas características principais a serem abordadas ao descrever dados.

- **Quantidade de dados.** Na maioria das técnicas de modelagem, há impasses associados ao tamanho dos dados. Grandes conjuntos de dados podem produzir modelos mais precisos, mas eles também podem aumentar o tempo de processamento. Considere se o uso de um subconjunto de dados é uma possibilidade. Ao tomar notas para o relatório final, certifique-se de incluir as estatísticas de tamanho para todos os conjuntos de dados e lembre-se de considerar o número de registros, bem como o de campos (atributos), ao descrever os dados.
- **Tipos de valor.** Os dados podem ter uma variedade de formatos, tais como **numérico**, **categórico** (sequência de caracteres) ou **booleano** (true/false). Prestar atenção ao tipo de valor pode prevenir problemas durante a modelagem posterior.
- **Esquemas de codificação.** Frequentemente, os valores no banco de dados são representações de características como sexo ou tipo de produto. Por exemplo, um conjunto de dados pode usar *M* e *F* para representar *homem* e *mulher*, enquanto outro pode usar os valores numéricos *1* e *2*. Anote qualquer esquema conflitante no relatório de dados.

Com esse conhecimento em mãos, você agora está pronto para escrever o relatório de descrição de dados e compartilhar suas descobertas com um público maior.

Exemplo de varejo eletrônico--Descrevendo dados

Um cenário de mineração na web usando o CRISP-DM

Há muitos registros e atributos a serem processados em um aplicativo de mineração na Web. Ainda que o varejista eletrônico realizando este projeto de mineração de dados tenha limitado o estudo inicial a aproximadamente 30.000 clientes que se registraram no site, ainda há milhões de registros nos logs da Web.

A maioria dos tipos de valores nessas fontes de dados é simbólica, sejam eles datas e horas, páginas da web acessadas ou respostas a perguntas de múltipla escolha do questionário do registro. Algumas dessas variáveis serão usadas para criar novas variáveis que são numéricas, como número de páginas da Web visitadas e o tempo gasto no website. As poucas variáveis numéricas existentes nas fontes de dados

incluem o número de cada produto solicitado, a quantia gasta durante uma compra e as especificações de peso e dimensão do produto do banco de dados do produto.

Há pouca sobreposição nos esquemas de codificação para as diversas fontes de dados, pois essas fontes de dados contêm atributos muito diferentes. As únicas variáveis que se sobrepõem são as "chaves", como IDs de clientes e códigos de produtos. Essas variáveis devem ter esquemas de codificação idênticos de fonte de dados a fontes de dados, do contrário seria impossível mesclar as fontes de dados. Será necessária alguma preparação de dados adicional para recodificar esses campos principais para mesclagem.

Escrevendo um relatório de descrição de dados

Para continuar de forma efetiva com seu projeto de mineração de dados, considere o valor da produção de um relatório preciso de descrição de dados usando as seguintes métricas:

Quantidade de dados

- Qual é o formato dos dados?
- Identifique o método usado para capturar os dados, por exemplo, ODBC.
- Qual é o tamanho do banco de dados (em números de linhas e colunas)?

Qualidade de dados

- Os dados incluem características relevantes a questões de negócios?
- Quais tipos de dados estão presentes (simbólico, numérico, etc.)?
- Você calculou as estatísticas básicas para os atributos-chave? Qual a percepção que isso forneceu nas questões de negócios?
- Você é capaz de priorizar atributos relevantes? Em caso negativo, os analistas de negócios estão disponíveis para fornecer maior percepção?

Explorando dados

Use esta fase do CRISP-DM para explorar os dados nas tabelas, nos gráficos e em outras ferramentas de visualização disponíveis no IBM SPSS Modeler. Tais análises podem ajudar a abordar a meta de mineração de dados construída durante a fase de entendimento dos negócios. Elas também podem ajudar a formular hipóteses e a dar forma às tarefas de transformação de dados que ocorrem durante a preparação de dados.

Exemplo de varejo eletrônico--Explorando dados

Um cenário de mineração na web usando o CRISP-DM

Embora o CRISP-DM sugira realizar uma exploração inicial neste ponto, a exploração de dados é difícil, se não impossível, em logs brutos da Web, como nosso varejista eletrônico descobriu. Normalmente, os dados do log da Web devem ser processados primeiro na fase de preparação de dados para produzir dados que possam ser explorados de forma significativa. Esse afastamento do CRISP-DM ressalta o fato de que o processo pode e deve ser customizado para suas necessidades específicas de mineração de dados. O CRISP-DM é cíclico e os mineradores de dados normalmente vão e vêm entre as fases.

Embora os logs da Web devam ser processados antes da exploração, as outras fontes de dados disponíveis para o varejista eletrônico são mais acessíveis à exploração. Usar o banco de dados de compras para exploração revela sumarizações interessantes sobre clientes, tais como quanto eles gastam, quantos itens eles adquirem por compra e de onde eles vêm. As sumarizações do banco de dados de clientes mostrarão a distribuição de respostas aos itens no questionário de registro.

A exploração também é útil para procurar erros nos dados. Embora a maior das fontes de dados sejam geradas automaticamente, as informações no banco de dados de produtos foram inseridas manualmente. Algumas sumarizações rápidas de dimensões de produtos listados ajudarão a descobrir erros de digitação como "monitor de 119 polegadas" (em vez de "19 polegadas").

Escrevendo um relatório de exploração de dados

À medida que criar gráficos e executar estatísticas sobre dados disponíveis, comece a formar hipóteses sobre como os dados podem atender as metas técnicas e de negócios.

Lista de tarefas

Tome notas de suas descobertas para a inclusão no relatório de exploração de dados. Certifique-se de responder às seguintes perguntas:

- Que tipo de hipótese você formulou sobre os dados?
- Quais atributos parecem promissores para maior análise?
- Suas explorações revelaram novas características sobre os dados?
- Como essas explorações alteraram sua hipótese inicial?
- Você pode identificar subconjuntos de dados para uso posterior?
- Dê uma outra olhada em suas metas de mineração de dados. Esta exploração alterou suas metas?

Verificando a qualidade dos dados

Os dados raramente são perfeitos. De fato, a maioria dos dados contém erros de codificação, valores omissos ou outros tipos de inconsistências que ocasionalmente tornam a análise delicada. Uma forma de evitar potenciais imprevistos é realizar uma análise de qualidade completa de dados disponíveis antes da modelagem.

As ferramentas de relatório no IBM SPSS Modeler (como Auditoria de Dados, Tabela e outros nós de saída) podem ajudá-lo com os seguintes tipos de problemas:

- **Dados omissos** incluem valores em branco ou codificados como uma não resposta (como *$null$*, *?* ou *999*).
- **Erros de dados** são normalmente erros tipográficos na entrada de dados.
- **Erros de medição** incluem dados que são inseridos corretamente, mas que se baseiam em um esquema de medição incorreto.
- **Inconsistências de codificação** normalmente envolvem unidades não padrão de medição ou de inconsistências de valor, como o uso de *H* e *homem* para sexo.
- **Metadados inválidos** incluem incompatibilidades entre o significado aparente de um campo e o significado indicado em um nome ou definição de campo.

Certifique-se de tomar notas sobre tais preocupações com a qualidade. Consulte o tópico "Escrevendo um relatório de qualidade de dados" na página 17 para obter informações adicionais.

Exemplo de varejo eletrônico--Verificando a qualidade dos dados

Um cenário de mineração na web usando o CRISP-DM

A verificação da qualidade dos dados é normalmente realizada no decorrer dos processos de descrição e exploração. Alguns dos problemas encontrados pelo varejista eletrônico incluem:

Dados omissos. Os dados omissos conhecidos incluem os questionários não respondidos por alguns dos usuários registrados. Sem as informações adicionais fornecidas pelo questionário, pode ser necessário que esses clientes sejam deixados de fora de alguns desses modelos subsequentes.

Erros de dados. A maioria das fontes de dados é gerada automaticamente, portanto isso não é uma grande preocupação. Erros tipográficos no banco de dados do produto podem ser encontrados durante o processo de exploração.

Erros de medição. A maior origem potencial para o erro de medição é o questionário. Se algum dos itens for mal recomendado ou mal formulado, ele poderá não fornecer as informações que o varejista eletrônico espera obter. Novamente, durante o processo de exploração, é importante prestar atenção especial a itens que têm uma distribuição incomum de respostas.

Escrevendo um relatório de qualidade de dados

Com base em sua exploração e verificação da qualidade de dados, você agora está pronto para preparar um relatório que orientará a próxima fase do CRISP-DM. Consulte o tópico “Verificando a qualidade dos dados” na página 16 para obter informações adicionais.

Lista de tarefas

Conforme discutido anteriormente, há diversos tipos de problemas de qualidade de dados. Antes de mover para a etapa seguinte, considere os seguintes interesses de qualidade e planeje uma solução. Documente todas as respostas no relatório de qualidade de dados.

- Você identificou atributos omissos e campos em branco? Em caso positivo, há algum significado por trás dos valores omissos?
- Há inconsistências ortográficas que possam causar problemas em mesclagens ou transformações mais recentes?
- Você explorou desvios para determinar se são "ruídos" ou se é um fenômeno que merece análise adicional?
- Você realizou uma verificação de plausibilidade dos valores? Tome notas sobre qualquer conflito aparente (tal como adolescentes com alto poder aquisitivo).
- Já considerou excluir dados que não têm nenhum impacto sobre suas hipóteses?
- Há arquivos armazenados em arquivos simples? Em caso positivo, os delimitadores são consistentes entre os arquivos? Cada registro contém o mesmo número de campos?

Pronto para o próximo passo?

Antes de preparar os dados para modelagem no IBM SPSS Modeler, considere os seguintes pontos:

Quanto você entende de dados?

- Todas as fontes de dados são claramente identificadas e acessadas? Você está ciente de algum problema ou restrição?
- Você identificou atributos-chave nos dados disponíveis?
- Esses atributos podem ajudá-lo a formular hipóteses?
- Você observou o tamanho de todas as fontes de dados?
- Você pode usar um subconjunto de dados onde apropriado?
- Você calculou as estatísticas básicas de cada atributo de interesse? Surgiu alguma informação significativa?
- Você usou gráficos exploratórios para obter uma maior percepção dos atributos-chave? Essa percepção reformulou alguma de suas hipóteses?
- Quais são os problemas de qualidade de dados deste projeto? Você tem um plano para abordar esses problemas?
- Os passos de preparação de dados estão claros? Por exemplo, você sabe quais origens de dados mesclar e quais atributos filtrar ou selecionar?

Agora que você está munido de entendimento de negócios e dados, é hora de usar o IBM SPSS Modeler para preparar seus dados para modelagem.

Capítulo 4. Preparação de Dados

Visão geral da preparação de dados

A preparação de dados é um dos aspectos mais importantes e frequentemente mais demorados da mineração de dados. De fato, estima-se que a preparação de dados normalmente consome 50-70% do tempo e esforço do projeto. Dedicar a energia adequada às primeiras fases de entendimento de negócios e entendimento de dados pode minimizar esse gasto adicional, mas ainda pode ser necessário empregar muito esforço na preparação e no empacotamento de dados para mineração.

Dependendo de sua organização e metas, a preparação de dados normalmente envolve as seguintes tarefas:

- Mesclar conjuntos e/ou registros de dados
- Selecionar um subconjunto de amostra de dados
- Agregar registros
- Derivar novos atributos
- Classificar dados para modelagem
- Remover ou substituir valores em branco ou omissos
- Dividir em conjuntos de dados de treinamento e de teste

Selecionando dados

Baseado na coleta inicial de dados conduzida na fase anterior do CRISP-DM, você está pronto para começar a selecionar os dados relevantes a suas metas de mineração de dados. Normalmente, há duas formas de selecionar dados:

- **Selecionar itens (linhas)** envolve tomada de decisões como quais contas, produtos ou clientes devem ser incluídos.
- **Selecionar atributos ou características (colunas)** envolve tomar decisões sobre o uso de características como quantia de transação ou renda doméstica.

Exemplo de varejo eletrônico--Selecionando dados

Um cenário de mineração na web usando o CRISP-DM

Muitas das decisões do varejista eletrônico sobre quais dados selecionar já foram tomadas em fases anteriores do processo de mineração de dados.

Selecionar itens. O estudo inicial será limitado a (aproximadamente) 30.000 clientes que se registraram no site, portanto os filtros precisam ser configurados para excluir compras e logs da Web de clientes não registrados. Outros filtros devem ser estabelecidos para remover chamadas de arquivos de imagem e outros entradas não informativas nos logs da Web.

Selecionar atributos. O banco de dados de compras conterá informações confidenciais sobre os clientes do varejista eletrônico, portanto é importante filtrar os atributos como nome, endereço, número do telefone e cartões de crédito do cliente.

Incluindo ou excluindo dados

À medida que você decide sobre subconjuntos de dados a serem incluídos ou excluídos, certifique-se de documentar a lógica por trás de suas decisões.

Perguntas a considerar

- Um dado atributo é relevante para suas metas de mineração de dados?

- A qualidade de um conjunto de dados ou um atributo específico impede a validade de seus resultados?
- Você pode recuperar os dados?
- Há alguma restrição sobre o uso de campos específicos como *sexo* ou *raça*?

Suas decisões aqui são diferentes daquelas das hipóteses formuladas na fase de entendimento dos dados? Em caso positivo, certifique-se de documentar sua argumentação no relatório do projeto.

Limpando os dados

Limpar os dados envolve olhar os problemas mais atentamente nos dados que você escolheu incluir para análise. Há diversas formas de limpar dados usando os nós de Operações de Registro e Campo no IBM SPSS Modeler.

Tabela 2. Limpando dados

Problema nos dados	**Solução possível**
Dados omissos	Exclua linhas ou características. Ou preencha os espaços em brancos por um valor estimado.
Erros de dados	Use a lógica para descobrir manualmente os erros e substitua-os. Ou exclua as características.
Inconsistências de codificação	Decida sobre um único esquema de codificação, então converta e substitua os valores.
Metadados omissos ou inválidos	Examine manualmente os campos suspeitos e rastreie o significado correto.

O Relatório de Qualidade de Dados preparado durante a fase de entendimento de dados contém detalhes sobre os tipos de problemas específicos dos seus dados. É possível usá-los como um ponto de partida para a manipulação de dados no IBM SPSS Modeler.

Exemplo de varejo eletrônico--Limpando dados

Um cenário de mineração na web usando o CRISP-DM

O varejista eletrônico usa o processo de limpeza de dados para abordar os problemas anotados no relatório de qualidade de dados.

Dados omissos. É provável que os clientes que não concluíram o questionário on-line tenham de ser deixados de lado de alguns modelos posteriormente. Pode-se requerer novamente que esses clientes preencham o questionário, mas isso exigirá tempo e dinheiro que o varejista eletrônico não pode se dar ao luxo de gastar. O que o varejista eletrônico pode fazer é modelar as diferenças de compras entre os clientes que respondem e não respondem ao questionário. Se esses dois conjuntos de clientes têm hábitos de compras semelhantes, os questionários omissos são menos preocupantes.

Erros de dados. Os erros encontrados durante o processo de exploração podem ser corrigidos aqui. No entanto, na maior parte dos casos, a entrada de dados apropriados é impingida no website antes de o cliente enviar uma página para o banco de dados de backend.

Erros de medição. Os itens redigidos de forma insatisfatória no questionário pode afetar muito a qualidade dos dados. Assim como com questionários omissos, esse é um problema difícil, pois pode não haver tempo ou dinheiro disponível para coletar respostas para uma nova pergunta de substituição. Para os itens problemáticos, a melhor solução pode ser voltar ao processo de seleção e filtrar esses itens de análises adicionais.

Escrevendo um relatório de limpeza de dados

Relatar os esforços de limpeza de dados é essencial para rastrear alterações nos dados. Os projetos de mineração de dados futuros irão se beneficiar de ter os detalhes de seu trabalho prontamente disponíveis.

Lista de tarefas

É uma boa ideia considerar as seguintes perguntas ao escrever o relatório:

- Quais tipos de ruído ocorreram nos dados?
- Quais abordagens você usou para remover o ruído? Quais técnicas foram bem-sucedidas?
- Há algum caso ou atributo que não pôde ser recuperado? Certifique-se de anotar os dados excluídos devido ao ruído.

Construindo novos dados

É frequente o caso de precisar construir novos dados. Por exemplo, pode ser útil criar uma nova coluna que sinalize a compra de uma garantia estendida para cada transação. Neste campo, *purchased_warranty* pode ser facilmente gerado usando um nó Configurar como Flag no IBM SPSS Modeler.

Há duas formas de construir novos dados:

- Derivando atributos (colunas ou características)
- Gerando registros (linhas)

O IBM SPSS Modeler oferece varias maneiras de construir dados usando os nós de Operações de Registro e Campo.

Exemplo de varejo eletrônico--Construindo dados

Um cenário de mineração na web usando o CRISP-DM

O processamento de logs da Web pode criar diversos novos atributos. Para os eventos registrados nos logs, o varejista eletrônico desejará criar registros de data e hora, identificar visitantes e sessões e anotar a página acessada e o tipo de atividade que o evento representa. Algumas dessas variáveis serão usadas para criar mais atributos, como o tempo entre os eventos em uma sessão.

Outros atributos podem ser criados como um resultado de uma mesclagem ou de outra reestruturação de dados. Por exemplo, quando os logs da Web de evento por linha forem "reunidos" para que cada linha seja uma sessão, serão criados novos atributos que gravam o número total de ações, o tempo total gasto e o total de compras feito durante a sessão. Quando os logs da Web forem mesclados com o banco de dados do cliente para que cada linha seja um cliente, serão criados novos atributos registrando o número de sessões, o número total de ações o tempo total gasto e total de compras feitas por cada cliente.

Depois de construir novos dados, o varejista eletrônico passa por um processo de exploração para se certificar de que a criação de dados foi executada corretamente.

Derivando atributos

No IBM SPSS Modeler, é possível usar os seguintes nós de Operações de Campo para derivar novos atributos:

- Crie novos campos derivados a partir daqueles existentes usando um **nó Derivar**.
- Crie um campo de flag usando um **nó Configurar como Flag**.

Lista de tarefas

- Considere os requisitos de dados para modelagem ao derivar atributos. O algoritmo de modelagem espera um tipo específico de dados, como o numérico? Em caso positivo, execute as transformações necessárias.
- Os dados precisam ser normalizados antes da modelagem?
- Os atributos ausentes podem ser construídos usando a agregação, a média ou a indução?
- Com base em seu conhecimento de segundo plano, há fatos importantes (como o tempo gasto no website) que podem ser derivados de campos existentes?

Integrando dados

Não é incomum ter diversas fontes de dados para o mesmo conjunto de questões de negócios. Por exemplo, você pode ter acesso a dados de empréstimo hipotecário, bem como ter adquirido dados demográficos para o mesmo conjunto de clientes. Se esses conjuntos de dados contiverem o mesmo identificador exclusivo (como o número de previdência social), será possível mesclá-los no IBM SPSS Modeler usando este campo-chave.

Há dois métodos básicos de integração de dados:

- **Mesclar** dados envolve mesclar dois conjuntos de dados com registros semelhantes, mas atributos diferentes. Os dados são mesclados usando o mesmo identificador de chave para cada registro (como o ID do cliente). Os dados resultantes aumentam em colunas ou características.
- **Anexar** dados envolve integrar dois ou mais conjuntos de dados com atributos semelhantes, mas com registros diferentes. Os dados são integrados com base em campos semelhantes (como nome do produto ou duração do contrato).

Exemplo de varejo eletrônico--Integrando dados

Um cenário de mineração na web usando o CRISP-DM

Com diversas fontes de dados, há diversas formas diferentes nas quais o varejista eletrônico pode integrar os dados:

- **Incluindo atributos de cliente e produto nos dados do evento.** Para modelar eventos de log da Web usando atributos de outros bancos de dados, cada ID de cliente, número de produto e número de ordem de compra associados a cada evento devem ser corretamente identificados e os atributos correspondentes devem ser mesclados nos logs da Web processados. Observe que o arquivo mesclado replica as informações de cliente e produto toda vez que um cliente ou produto é associado a um evento.
- **Incluindo informações de compra e log da Web nos dados do cliente.** Para modelar o valor de um cliente, suas informações de compras e sessão devem ser escolhidas nos bancos de dados apropriados, totalizadas e mescladas com o banco de dados do cliente. Isso envolve a criação de novos atributos, conforme discutido no processo de construção de dados.

Depois de integrar os bancos de dados, o varejista eletrônico passa por um processo de exploração para se certificar de que a mesclagem de dados foi executada corretamente.

Tarefas de integração

A integração de dados pode se tornar complexa se você não empregou tempo suficiente desenvolvendo e entendendo seus dados. Reflita um pouco sobre os itens e atributos que parecerem mais relevantes para as metas de mineração de dados e, então, comece a integração de dados.

Lista de tarefas

- Usar os nós Mesclar ou Anexar no IBM SPSS Modeler, integra os conjuntos de dados considerados úteis para modelagem.
- Considere salvar a saída resultante antes de continuar com a modelagem.
- Após a mesclagem, os dados podem ser simplificados ao **agregar** valores. A agregação significa que os novos valores são calculados ao sumarizar as informações de diversos registros e/ou tarefas.
- Também pode ser necessário gerar novos registros (tais como a dedução da média de diversos anos de declarações de imposto de renda combinadas).

Formatação de dados

Como uma etapa final antes da construção do modelo, é útil verificar se determinadas técnicas requerem uma ordem ou um formato específico para os dados. Por exemplo, não é incomum que um algoritmo de sequência requeira que os dados sejam ordenados previamente antes da execução do modelo. Mesmo

que o modelo possa executar a ordenação para você, o uso de um nó Ordenar antes da modelagem poderá economizar tempo de processamento.

Lista de tarefas

Considere as seguintes perguntas ao formatar os dados:

- Quais modelos você planeja usar?
- Esses modelos requerem um formato ou uma ordem de dados específico?

Se mudanças forem recomendadas, as ferramentas de processamento no IBM SPSS Modeler poderão ajudá-lo a aplicar a manipulação necessária de dados.

Pronto para a modelagem?

Antes de construir os modelos no IBM SPSS Modeler, certifique-se de ter respondido as seguintes perguntas.

- Todos os dados estão acessíveis a partir do IBM SPSS Modeler?
- Com base em sua exploração e entendimento iniciais, você conseguiu selecionar subconjuntos relevantes de dados?
- Você limpou os dados de forma efetiva ou removeu os dados irrecuperáveis? Documente qualquer decisão no relatório final.
- Os diversos conjuntos de dados estão integrados adequadamente? Ocorreu algum problema de mesclagem que deva ser documentado?
- Você pesquisou os requisitos das ferramentas de modelagem que planeja usar?
- Há problemas de formatação que possam ser abordados antes da modelagem? Isso inclui questões de formatação necessária, bem como tarefas que possam reduzir o tempo de modelagem.

Se você puder responder as perguntas acima, então estará pronto para o ponto crucial da mineração de dados, a modelagem.

Capítulo 5. Modelagem

Visão Geral de Modelagem

Este é o ponto em que seu grande esforço começa a compensar. Os dados que você levou tempo preparando são levados às ferramentas de análise no IBM SPSS Modeler e os resultados começam a esclarecer os problemas de negócios apresentados durante o Entendimento dos negócios.

A modelagem normalmente é realizada em diversas iterações. Normalmente, os mineradores de dados executam diversos modelos usando os parâmetros padrão e, então, ajustam os parâmetros ou revertem para a fase de preparação de dados para as manipulações requeridas por seu modelo de preferência. É raro em uma organização que a pergunta da mineração de dados seja respondida satisfatoriamente com um único modelo e uma única execução. É isso que torna a mineração de dados tão interessante; há várias formas de olhar para um determinado problema e o IBM SPSS Modeler oferece uma ampla variedade de ferramentas para ajudá-lo a fazer isso.

Selecionando técnicas de modelagem

Embora você já tenha alguma ideia sobre quais tipos de modelagem são mais apropriados para as necessidades de sua organização, agora é hora de tomar algumas decisões firmes sobre quais usar. A determinação do modelo mais apropriado normalmente irá se basear nas seguintes considerações:

- **Os tipos de dados disponíveis para mineração.** Por exemplo, os campos de interesse são categóricos (simbólicos)?
- **Suas metas de mineração de dados.** Você deseja simplesmente obter a percepção de armazenamentos de dados transacionais e revelar padrões de compras interessantes? Ou você precisa produzir um escore indicando, por exemplo, a propensão à inadimplência em um empréstimo estudantil?
- **Requisitos específicos de modelagem.** O modelo requer um tamanho ou tipo específico de dados? Você precisa de um modelo com resultados facilmente apresentáveis?

Para obter mais informações sobre os tipos de modelos no IBM SPSS Modeler e seus requisitos, consulte a documentação do IBM SPSS Modeler ou a Ajuda on-line.

Exemplo de varejo eletrônico--Técnicas de modelagem

As técnicas de modelagem empregadas pelo varejista eletrônico são conduzidas pelas metas de mineração de dados da empresa:

Recomendações melhoradas. Da forma mais simples, isso envolve ordens de compra de armazenamento em cluster para determinar quais produtos são comprados juntos com maior frequência. Os dados do cliente, e até os registros de visita, podem ser incluídos para a obtenção de melhores resultados. As técnicas de armazenamento em cluster de rede Kohonen são apropriados para esse tipo de modelagem. Posteriormente, os clusters podem ser modelados usando um conjunto de regras do C5.0 para determinar quais recomendações são as mais apropriadas em qualquer ponto durante a visita de um cliente.

Navegação de site melhorada. Por enquanto, o varejista eletrônico irá se concentrar na identificação de páginas que são frequentemente usadas, mas que requerem diversos cliques para que o usuário as encontrem. Isso requer aplicar um algoritmo de sequenciamento nos logs da Web a fim de gerar os "caminhos exclusivos" que os clientes podem pegar no website e, então, procurar especificamente por sessões com muitas páginas a serem visitadas sem (ou antes) que uma medida seja tomada. Posteriormente, em uma análise mais profunda, as técnicas de armazenamento em cluster poderão ser usadas para identificar diferentes "tipos" de visitas e de visitantes e o conteúdo do site poderá ser organizado e apresentado de acordo com o tipo.

Escolhendo as técnicas corretas de modelagem

Muitas técnicas de modelagem estão disponíveis no IBM SPSS Modeler. Frequentemente, os mineradores de dados usam mais de uma para abordar o problema sob diversas direções.

Lista de tarefas

Ao decidir em quais modelos usar, considere se os seguintes problemas têm um impacto em suas opções:

- O modelo requer que os dados sejam divididos em conjuntos de teste e de treinamento?
- Você tem dados suficientes para produzir resultados confiáveis para um determinado modelo?
- O modelo requer um certo nível de qualidade de dados? Você pode satisfazer esse nível com os dados atuais?
- Os seus dados são do tipo apropriado para um modelo específico? Em caso negativo, você pode fazer as conversões necessárias usando nós de manipulação de dados?

Para obter mais informações sobre os tipos de modelos no IBM SPSS Modeler e seus requisitos, consulte a documentação do IBM SPSS Modeler ou a Ajuda on-line.

Suposições de modelagem

À medida que você começa a limitar as ferramentas de modelagem de sua preferência, tome notas sobre o processo de tomada de decisão. Documente qualquer suposição de dados, bem como qualquer manipulação de dados feita para atender os requisitos do sistema.

Por exemplo, os nós de Regressão Logística e Rede Neural requerem que os tipos de dados sejam totalmente **instanciados** (os tipos de dados são conhecidos) antes da execução. Isso significa que será necessário incluir um nó de Tipo no fluxo e executá-lo para que os dados passem por ele, antes de construir e executar um modelo. Da mesma forma, os modelos preditivos, como o C5.0, podem se beneficiar do reequilíbrio dos dados ao prever regras para eventos raros. Ao fazer esse tipo de previsão, você normalmente obtém resultados melhores inserindo um nó de Balanceamento no fluxo e alimentando o subconjunto mais balanceado no modelo.

Certifique-se de documentar esses tipos de decisões.

Gerando um design de teste

Como uma etapa final antes da construção real do modelo, é necessário tirar um momento para considerar novamente como os resultados do modelo serão testados. Há duas partes para gerar um design de teste abrangente:

- Descrever os critérios de "excelência" de um modelo
- Definir os dados nos quais esses critérios serão testados

A **excelência** pode ser medida de diversas formas. Para os modelos supervisionados, como o C5.0 e o C&R Tree, as medições de excelência normalmente estima a taxa de erro de um modelo específico. Para modelos não supervisionados, como as redes de cluster Kohonen, as medições podem incluir critérios como a facilidade de interpretação, a implementação ou o tempo de processamento.

Lembre-se: a construção de modelo é um processo interativo. Isso significa que você normalmente testará os resultados de diversos modelos antes de decidir sobre aqueles a serem usados e implementados.

Escrevendo um design de teste

O design de teste é uma descrição dos passos que serão executados para testar os modelos produzidos. Como a modelagem é um processo iterativo, é importante saber quando parar o ajuste de parâmetros e tentar outro método ou modelo.

Lista de tarefas

Ao criar um design de teste, considere as seguintes questões:

- Quais dados serão usados para testar os modelos? Você particionou os dados em conjuntos de treino/teste? (Essa é uma abordagem normalmente usada na modelagem.)
- Como você pode medir o sucesso de modelos supervisionados (como o C5.0)?
- Como você pode medir o sucesso de modelos não supervisionados (como redes de cluster Kohonen)?
- Quantas vezes pretende executar novamente um modelo com configurações ajustadas, antes de tentar outro tipo de modelo?

Exemplo de varejo eletrônico--Testar design

Um cenário de mineração na web usando o CRISP-DM

Os critérios pelos quais os modelos são avaliados dependem dos modelos sob consideração e das metas de mineração de dados:

Recomendações melhoradas. Até que as recomendações melhoradas estejam presentes em tempo real para os clientes, não há nenhuma forma puramente objetiva de avaliá-los. Entretanto, o varejista eletrônico pode requerer que as regras que geram as recomendações sejam simples o suficiente para fazer sentido a partir de uma perspectiva do negócio. Da mesma forma, as regras devem ser suficientemente complexas para gerar recomendações diferentes para clientes e sessões diferentes.

Navegação de site melhorada. Dada a evidência de quais páginas os clientes acessam no website, o varejista eletrônico pode objetivamente avaliar o design atualizado do site em termos da facilidade de acesso a páginas importantes. Entretanto, da mesma forma que as recomendações, é difícil avaliar antecipadamente como os clientes irão se ajustar ao site reorganizado. Se o tempo e s finanças permitirem, algum teste de usabilidade deve estar preparado.

Construindo os modelos

Neste ponto, você deve estar bem preparado para construir os modelos que você passou tanto tempo levando em consideração. Dê-se tempo e espaço para experimentar alguns modelos diferentes antes e tomar a decisão final. A maioria dos mineradores de dados normalmente constroem diversos modelos e comparam os resultados antes de implementá-los ou integrá-los.

Para controlar seu progresso com uma variedade de modelos, certifique-se de manter notas sobre as configurações e os dados usados para cada modelo. Isso irá ajudá-lo a discutir os resultados com outros e rastrear seus passos novamente, se necessário. No final do processo de construção do modelo, você terá três informações a serem usadas nas decisões de mineração de dados:

- **Configurações de parâmetros** incluem as notas feitas sobre os parâmetros que produzem os melhores resultados.
- Os **modelos** reais produzidos.
- **Descrições de resultados do modelo**, incluindo o desempenho e os problemas de dados ocorridos durante a execução do modelo e a exploração de seus resultados.

Exemplo de varejo eletrônico--Construção de modelo

Um cenário de mineração na web usando o CRISP-DM

Recomendações melhoradas. As clusterizações são produzidas para níveis variáveis de integração de dados, iniciando apenas com o banco de dados de compra e, em seguida, incluindo as informações sobre o cliente relacionado e a sessão. Para cada nível de integração, as clusterizações são produzidas sob configurações variáveis de parâmetros para os algoritmos de rede em duas etapas e Kohonen. Para cada uma dessas clusterizações, alguns conjuntos de regras C5.0 são gerados com diferentes configurações de parâmetros.

Navegação de site melhorada. O nó de modelagem Sequência é usado para gerar caminhos do cliente. O algoritmo permite a especificação de um critério de suporte mínimo, o qual é útil para se concentrar nos caminhos mais comuns do cliente. São testadas diversas configurações para os parâmetros.

Configurações de parâmetro

A maioria das técnicas de modelagem têm uma variedade de parâmetros ou configurações que podem ser ajustados para controlar o processo de modelagem. Por exemplo, as árvores de decisão podem ser controladas ajustando a profundidade da árvore, as divisões e diversas outras configurações. Normalmente, a maioria das pessoas constrói um modelo usando primeiro as opções padrão e, então, refina os parâmetros durante sessões subsequentes.

Assim que determinar os parâmetros que produzem os resultados mais precisos, certifique-se de salvar os nós do fluxo e do modelo gerado. Além disso, tomar notas das configurações ideais pode ajudá-lo quando decidir automatizar ou reconstruir o modelo com novos dados.

Executando os modelos

No IBM SPSS Modeler, a execução de modelos é uma tarefa direta. Assim que tiver inserido o nó do modelo no fluxo e editado qualquer parâmetro, simplesmente execute o modelo para produzir resultados visualizáveis. Os resultados aparece no navegador Modelos Gerados, no lado direito da área de trabalho. É possível clicar com o botão direito em um modelo para navegar nos resultados. Para a maioria dos modelos, é possível inserir o modelo gerado no fluxo para melhor avaliar e implementar os resultados. Os modelos também podem ser salvos no IBM SPSS Modeler para uma fácil reutilização.

Descrição da modelo

Ao examinar os resultados de um modelo, certifique-se de tomar notas sobre sua experiência com a modelagem. É possível armazenar essas notas no próprio modelo usando a caixa de diálogo de anotações ou a ferramenta do projeto.

Lista de tarefas

Para cada modelo, registre as informações como:

- É possível tirar conclusões significativas a partir desse modelo?
- Foram reveladas novas percepções ou padrões incomuns pelo modelo?
- Ocorreram problemas de execução do modelo? O tempo de processamento foi razoável?
- O modelo apresentou dificuldades com os problemas de qualidade de dados, tais como um alto número de valores omissos?
- Ocorreu alguma inconsistência de cálculo digna de atenção?

Avaliando o modelo

Agora que você já tem um conjunto inicial de modelos, observe-os mais atentamente para determinar quais são precisos ou eficazes o suficiente para serem finais. Final pode significar várias coisas, como "pronto para implementação" ou "ilustrando padrões interessantes". Consultar o plano de teste criado anteriormente pode ajudá-lo a fazer essa avaliação a partir do ponto de vista de sua organização.

Avaliação abrangente de modelo

Para cada modelo sob consideração, é uma boa ideia fazer uma avaliação metódica com base nos critérios gerados em seu plano de teste. Aqui é onde você pode incluir o modelo gerado no fluxo e usar os gráficos de avaliação ou os nós de análise para analisar a eficácia dos resultados. Você também deve considerar se os resultados fazem sentido lógico ou se eles são simplistas demais para os seus objetivos de negócio (por exemplo, uma sequência que revela compras como o vinho> vinho> vinho).

Assim que tiver feito uma avaliação, classifique os modelos na ordem com base nos critérios objetivos (precisão do modelo) e subjetivos (facilidade de uso ou interpretação de resultados).

Lista de tarefas

- O uso de ferramentas de mineração de dados no IBM SPSS Modeler, como gráficos de avaliação, nós de análise ou gráficos de validação cruzada, avalia os resultados de seu modelo.
- Realize uma revisão dos resultados com base em seu entendimento de problemas de negócios. Consulte analistas de dados ou outros especialistas que possam ter uma percepção da relevância de resultados específicos.
- Considere se os resultados do modelo são facilmente implementáveis. Sua organização requer que os resultados sejam implementados na Web ou que sejam enviados de volta para o data warehouse?
- Analise o impacto dos resultados sobre seus critérios de sucesso. Eles atendem as metas estabelecidas durante a fase de entendimento dos negócios?

Se você foi capaz de abordar os problemas acima com êxito e acredita que os modelos atuais atendem suas metas, é hora de passar para uma avaliação mais detalhada dos modelos e para uma implementação final. Do contrário, pegue aquilo que aprendeu e execute novamente os modelos com configurações de parâmetro ajustado.

Exemplo de varejo eletrônico--Avaliação de modelo

Um cenário de mineração na web usando o CRISP-DM

Recomendações melhoradas. Uma das redes Kohonen e uma clusterização em duas etapas oferecem resultados razoáveis e o varejista eletrônico tem dificuldades para escolher entre elas. Com o tempo, a empresa espera usar ambas, aceitando as recomendações de que as duas técnicas combinam e estudando detalhadamente as situações nas quais elas diferem. Com um pouco de esforço e conhecimento aplicado de negócios, o varejista eletrônico pode desenvolver outras regras para resolver as diferenças entre as duas técnicas.

O varejista eletrônico também descobre que os resultados que incluem as informações da sessão são surpreendentemente bons. Há evidências que sugerem que as recomendações podem ser ligadas à navegação do site. Um conjunto de regras, definindo para onde o cliente provavelmente irá a seguir, pode ser usado em tempo real para afetar o conteúdo do site diretamente enquanto o cliente está navegando.

Navegação de site melhorada. O modelo Sequência fornece ao varejista eletrônico um alto nível de confiança de que determinados caminhos do cliente podem ser previstos, produzindo resultados que sugerem um número gerenciável de mudanças no design do site.

Mantendo o controle dos parâmetros revisados

Com base naquilo que você aprendeu durante a avaliação do modelo, é hora de dar uma outra olhada nos modelos. Aqui você tem duas opções:

- Ajustar os parâmetros de modelos existentes.
- Escolher um modelo diferente para abordar seu problema de mineração de dados.

Em ambos os casos, você voltará para a tarefa de construção de modelos e iterará até que os resultados sejam bem-sucedidos. Não se preocupe com a repetição deste passo. É extremamente comum que os mineradores de dados avaliem e executem novamente os modelos diversas vezes, até encontrar um que atenda a suas necessidades. Esse é um bom argumento para construir diversos modelos de uma vez e comparar os resultados antes de ajustar os parâmetros para cada um.

Pronto para o próximo passo?

Antes de continuar para uma avaliação final dos modelos, considere se a avaliação inicial foi completa o suficiente.

Lista de tarefas

- Você está apto a compreender os resultados dos modelos?
- Os resultados do modelo fazem sentido para você sob a perspectiva puramente lógica? Existem inconsistências aparentes que precisam de maior exploração?
- A partir de uma olhada inicial, os resultados parecem abordar as questões de negócios da organização?

- Você usou nós de análise e gráficos de ganhos para comparar e avaliar a precisão do modelo?
- Você explorou mais de um tipo de modelo e comparou os resultados?
- Os resultados de seu modelo são implementáveis?

Se os resultados de sua modelagem de dados parecerem precisos e relevantes, está no momento de realizar uma avaliação completa antes da implementação final.

Capítulo 6. Avaliação

Visão geral da avaliação

Neste ponto, você completou a maior parte de seu projeto de mineração de dados. Você também determinou, na fase de Modelagem, que os modelos construídos são tecnicamente corretos e efetivos de acordo com os **critérios de sucesso da mineração de dados** definidos anteriormente.

Entretanto, antes de continuar, é necessário avaliar os resultados de seus esforços usando os **critérios de sucesso dos negócios** estabelecidos no início do projeto. Essa é a chave para garantir que sua organização pode usar os resultados obtidos. Dois tipos de resultados são produzidos pela mineração de dados:

- Os **modelos** finais selecionados na fase anterior do CRISP-DM.
- Quaisquer conclusões ou inferências tiradas dos próprios modelos, bem como do processo de mineração de dados. Elas são conhecidas como **descobertas**.

Avaliando os resultados

Nesta etapa, você formaliza sua avaliação sobre se os resultados do projeto atendem ou não os critérios de sucesso dos negócios. Esta etapa requer um claro entendimento das metas de negócios determinadas, portanto tenha certeza de incluir os principais tomadores de decisão na avaliação do projeto.

Lista de tarefas

Primeiro, é necessário documentar sua avaliação de se os resultados da mineração de dados atendem ou não os critérios de sucesso dos negócios. Considere as seguintes questões em seu relatório:

- Os seus resultados estão indicados claramente e de uma forma que possam ser facilmente apresentados?
- Há descobertas particularmente novas ou exclusivas que devem ser destacadas?
- Você pode ranquear os modelos e as descobertas de acordo com sua aplicabilidade às metas dos negócios?
- Em geral, como esses resultados respondem às metas de negócios de sua organização?
- Quais outras perguntas seus resultados levantaram? Como você pode expressar essas perguntas em termos de negócios?

Depois de ter avaliado os resultados, compile uma lista de modelos aprovados para inclusão no relatório final. Essa lista deve incluir modelos que atendam à mineração de dados e às metas de negócios de sua organização.

Exemplo de varejo eletrônico--Avaliando resultados

Um cenário de mineração na web usando o CRISP-DM

Os resultados gerais da primeira experiência do varejista eletrônico com a mineração de dados são bem fáceis de serem transmitidos a partir de uma perspectiva do negócio: o estudo produziu aquilo que se espera que sejam melhores recomendações do produto e um design de site aprimorado. O design aprimorado do site baseia-se nas sequências de navegação do cliente, as quais mostram os recursos do site que os clientes desejam, mas que requerem vários passos para que sejam atingidos. A evidência de que as recomendações do produto são melhores é mais difícil de transferir, pois as regras de decisão podem se tornar complicadas. Para produzir o relatório final, os analistas tentarão identificar algumas tendências gerais nos conjuntos de regras que podem ser explicadas mais facilmente.

Ranqueando os modelos. Como vários dos modelos iniciais pareceram fazer sentido nos negócios, o ranqueamento nesse grupo se baseou nos critérios estatísticos, na facilidade de interpretação e na diversidade. Assim, o modelo forneceu diferentes recomendações para diferentes situações.

Novas perguntas. A pergunta mais importante que surgiu do estudo é: Como o varejista eletrônico pode saber mais sobre seus clientes? As informações no banco de dados de clientes desempenha um importante papel na formação dos clusters para recomendações . Embora regras especiais estejam disponíveis para fazer recomendações a clientes cujas informações estejam omissas, as recomendações são mais gerais por natureza do que aquelas feitas a clientes registrados.

Processo de revisão

As metodologias efetivas normalmente incluem tempo para reflexão sobre os êxitos e as fraquezas do processo recém-concluído. A mineração de dados não é diferente. Parte do CRISP-DM é aprender com sua experiência, para que os futuros projetos de mineração de dados sejam mais efetivos.

Lista de tarefas

Primeiro, deve-se resumir as atividades e decisões para cada fase, incluindo etapas de preparação de dados, construção de modelo, etc. Em seguida, para cada fase, considere as seguintes perguntas e faça sugestões de melhoria:

- Este estágio contribui para o valor dos resultados finais?
- Há formas de aperfeiçoar ou melhorar este estágio ou operação específico?
- Quais foram as falhas ou erros desta fase? Como poderão ser evitados na próxima vez?
- Você se deparou com alguma dificuldade, como modelos específicos que se mostraram infrutíferos? Há formas de prever tais dificuldades para que os esforços possam ser direcionados de forma mais produtiva?
- Houve alguma surpresa (boa ou ruim) durante esta fase? Em retrospectiva, há alguma forma óbvia de prever tais ocorrências?
- Há decisões ou estratégias alternativas que poderiam ter sido usadas em uma determinada fase? Anote tais alternativas para futuros projetos de mineração de dados.

Exemplo de varejo eletrônico--Revisar relatório

Um cenário de mineração na web usando o CRISP-DM

Como resultado da revisão do processo do projeto de mineração de dados inicial, o varejista eletrônico desenvolveu uma maior apreciação das inter-relações entre os passos no processo. Inicialmente relutante para "retroceder" no processo do CRISP-DM, o varejista eletrônico agora vê que a natureza cíclica do processo aumenta seu poder. A revisão do processo também levou o varejista eletrônico a entender que:

- Um retorno ao processo de exploração é sempre justificado quando algo incomum aparece em outra fase do processo do CRISP-DM.
- A preparação de dados, especialmente de logs da Web, requer paciência, visto que isso pode levar muito tempo.
- É essencial manter-se focado nos problemas de negócios em mãos, pois assim que os dados estão prontos para análise, é muito fácil começar a construir modelos sem levar em consideração uma imagem mais ampla.
- Assim que a fase de modelagem termina, o entendimento dos negócios é ainda mais importante na decisão de como implementar resultados e determinar quais outros estudos são justificados.

Determinando os próximos passos

Até agora, você produziu resultados, avaliou suas experiências de mineração de dados e pode estar se perguntando: **O que fazer a seguir?** Esta fase o ajuda a responder essa pergunta à luz de suas metas de negócios para a mineração de dados. Neste ponto, você tem basicamente duas opções:

- **Seguir para a fase de implementação.** A fase seguinte irá ajudá-lo a incorporar os resultados do modelo a seu processo de negócios e a produzir um relatório final. Mesmo que seus esforços na mineração de dados tenham sido malsucedidos, é necessário usar a fase de implementação do CRISP-DM para criar um relatório final para distribuição para o patrocinador do projeto.
- **Volte e refine ou substitua seus modelos.** Se você achar que seus resultados estão quase ideais, mas não totalmente, considere outra rodada de modelagem. Você pode pegar o que aprendeu nesta fase e usar para refinar os modelos e produzir melhores resultados.

Neste ponto, sua decisão envolve a precisão e a relevância dos resultados da modelagem. Se os resultados abordarem sua mineração de dados e suas metas de negócios, então você estará pronto para a fase de implementação. Qualquer que seja sua decisão, certifique-se de documentar o processo de avaliação detalhadamente.

Exemplo de varejo eletrônico--Etapas seguintes

Um cenário de mineração na web usando o CRISP-DM

O varejista eletrônico está bem confiante da precisão e da relevância dos resultados do projeto e, portanto, está prosseguindo para a fase de implementação.

Ao mesmo tempo, a equipe do projeto também está pronta para voltar e aumentar alguns dos modelos para que incluam as técnicas preditivas. Neste ponto, eles estão aguardando pela entrega dos relatórios finais e por uma luz verde dos tomadores de decisões.

Capítulo 7. Implementação

Visão geral da implementação

A implementação é o processo de uso de suas percepções para fazer melhorias em sua organização. Isso pode significar uma integração formal, tal como a implementação de um modelo do IBM SPSS Modeler que produz escores de perda de clientes que são então lidos em um data warehouse. Como alternativa, a implementação pode significar que você usa as percepções obtidas da mineração de dados para induzir mudanças em sua organização. Por exemplo, talvez você tenha descoberto padrões alarmantes em seus dados, indicando uma mudança no comportamento para clientes acima dos 30 anos. Esses resultados poderão não ser formalmente integrados a seus sistemas de informações, mas, sem dúvida, eles serão úteis para o planejamento e para as tomadas de decisões de marketing.

Em geral, a fase de implementação do CRISP-DM inclui dois tipos de atividades:

- Planejamento e monitoramento da implementação de resultados
- Conclusão de tarefas de finalização, tais como a produção de um relatório final e a realização de uma revisão do projeto

Dependendo dos requisitos de sua organização, poderá ser necessário concluir um ou ambos os passos.

Planejando para Implementação

Embora você possa estar ansioso para compartilhar os frutos de seus esforços na mineração de dados, tire um tempo para planejar uma implementação suave e abrangente dos resultados.

Lista de tarefas

- O primeiro passo é resumir os resultados, tanto modelos como descobertas. Isso o ajuda a determinar quais modelos podem ser integrados a seus sistemas de banco de dados e quais descobertas devem ser apresentadas a seus colegas.
- Para cada modelo implementável, crie um plano passo a passo para a implementação e integração com seus sistemas. Anote qualquer detalhe técnico, como requisitos do banco de dados para a saída do modelo. Por exemplo, talvez seu sistema requeira que a saída do modelo seja implementada em um formato delimitado por tabulação.
- Para cada descoberta conclusiva, crie um plano para disseminar essas informações para os desenvolvedores de estratégia.
- Há planos de implementação alternativos para ambos os tipos de resultados que valem ser mencionados?
- Considere como a implementação será monitorada. Por exemplo, como um modelo implementado usando o IBM SPSS Modeler Solution Publisher será ser atualizado? Como você decidirá quando um modelo não é mais aplicável?
- Identifique qualquer problema de implementação e planeje as contingências. Por exemplo, os tomadores de decisão podem querer mais informações sobre resultados de modelagem e podem requerer que forneça maiores detalhes técnicos.

Exemplo de varejo eletrônico--Planejamento de implementação

Um cenário de mineração na web usando o CRISP-DM

Uma implementação bem-sucedida dos resultados de mineração de dados do varejista eletrônico requer que as informações corretas atinjam as pessoas certas.

Tomadores de decisão. Os tomadores de decisões precisam ser informados das recomendações e mudanças propostas para o site e receber breves explicações de como essas mudanças ajudarão.

Presumindo que eles aceitem os resultados do estudo, as pessoas que implementarão as mudanças precisam ser notificadas.

Desenvolvedores da Web. As pessoas que mantêm o website terão de incorporar as novas recomendações e a organização do conteúdo do site. Informe-os sobre as mudanças que *poderão* ocorrer devido a estudos futuros, para que eles possam lançar as bases agora. Preparar a equipe para a rápida construção do site com base na análise de sequência em tempo real poderá ser útil posteriormente.

Especialistas em banco de dados. As pessoas que mantêm os bancos de dados de clientes, compras e produtos devem ser avisadas de como as informações dos bancos de dados estão sendo usadas e quais atributos podem ser incluídos nos bancos de dados em projetos futuros.

Acima de tudo, a equipe do projeto precisa manter contato com cada um desses grupos para coordenar a implementação de resultados e planejar os futuros projetos.

Planejando o monitoramento e a manutenção

Em uma implementação e integração completa de resultados de modelagem, seu trabalho de mineração de dados pode ser contínuo. Por exemplo, se um modelo for implementado para prever sequências de compras de cesta eletrônica, esse modelo provavelmente precisará ser avaliado periodicamente para garantir sua eficácia e fazer melhorias contínuas. Da mesma forma, um modelo implementado para aumentar a retenção do cliente entre clientes de alto valor provavelmente precisará ser ajustado assim que um nível específico de retenção for atingido. O modelo poderá, então, ser modificado e reutilizado para reter clientes em um nível mais baixo, mas ainda lucrativo, na pirâmide de valores.

Lista de tarefas

Tome nota dos seguintes problemas e certifique-se de incluí-los no relatar final.

- Para cada modelo ou descoberta, quais fatores ou influências (como valor de mercado o variação sazonal) precisam ser rastreados?
- Como a validade e a precisão de cada modelo podem ser medidas e monitoradas?
- Como você determinará quando um modelo "expirou"? Forneça condições específicas sobre limites ou mudanças esperadas nos dados, etc.
- O que acontecerá quando um modelo expirar? Você pode simplesmente reconstruir o modelo com dados mais novos ou fazer leves ajustes? Ou as mudanças serão disseminadas o suficiente para exigir um novo projeto de mineração de dados?
- Esse modelo pode ser usado para problemas de negócios semelhantes assim que tiver expirado? Este é o ponto em que uma boa documentação torna-se essencial para avaliar o propósito comercial de cada projeto de mineração de dados.

Exemplo de varejo eletrônico--Monitoramento e manutenção

Um cenário de mineração na web usando o CRISP-DM

A tarefa imediata para o monitoramento é determinar se a organização do novo site e as recomendações melhoradas realmente funcionam. Ou seja, os usuários podem seguir rotas mais diretas para as páginas que estão procurando? As vendas cruzadas de itens recomendados aumentaram? Após algumas semanas de monitoramento, o varejista eletrônico poderá determinar o sucesso do estudo.

O que pode ser manipulado automaticamente é a inclusão de novos usuários registrados. Quando os clientes se registram no site, os conjuntos de regras atuais podem ser aplicados a suas informações para determinar quais recomendações eles devem receber.

Decidir quando atualizar os conjuntos de regras para determinar recomendações é uma tarefa mais delicada. Atualizar os conjuntos de regras não é um processamento automático, pois a criação do cluster requer entrada manual referente à apropriabilidade de uma determinada solução de cluster.

Visto que projetos futuros geram modelos mais complexos, é quase certo que a necessidade e a quantidade de monitoramento aumentarão. Quando possível, o volume do monitoramento deve ser automático com relatórios planejados com regularidade disponíveis para revisão. Como alternativa, a

criação de modelos que fornecem predições rapidamente pode ser uma direção que a empresa gostaria de seguir. Isso requer uma maior sofisticação da equipe do que o primeiro projeto de mineração de dados.

Produzindo um relatório final

Escrever um relatório final não apenas liga as pontas soltas na documentação anterior, mas também pode ser usado para comunicar os seus resultados. Embora isso possa parecer direto, é importante apresentar seus resultados a diversas pessoas com um interesse nos resultados. Isso pode incluir administradores técnicos, que serão responsáveis pela implementação dos resultados da modelagem, bem como patrocinadores de marketing e gerenciamento, que tomarão suas decisões com base em seus resultados.

Lista de tarefas

Primeiro, considere o público de seu relatório. Eles são desenvolvedores técnicos ou gerentes focados no mercado? Poderá ser necessário criar relatórios separados para cada público, caso suas necessidades sejam incompatíveis. Em qualquer caso, o seu relatório deve incluir a maioria dos seguintes pontos:

- Um descrição completa dos problemas de negócios originais
- O processo usado para conduzir a mineração de dados
- Custos do projeto
- Observações sobre quaisquer desvios do plano original do projeto
- Um resumo de resultados da mineração de dados, tanto modelos como descobertas
- Uma visão geral do plano proposto de implementação
- Recomendações para um trabalho de mineração de dados mais detalhado, incluindo oportunidades interessantes descobertas durante a exploração e a modelagem

Preparando uma apresentação final

Além do relatório do projeto, também pode ser necessário apresentar as descobertas do projeto a uma equipe de patrocinadores ou a departamentos relacionados. Se for esse o caso, você pode usar grande parte das mesmas informações em seu relatório, mas apresentadas sob uma perspectiva mais ampla. Os diagramas e gráficos no IBM SPSS Modeler podem ser facilmente exportados para esse tipo de apresentação.

Exemplo de varejo eletrônico--Relatório final

Um cenário de mineração na web usando o CRISP-DM

O maior desvio do plano original do projeto também é uma oportunidade interessante para um maior trabalho de mineração de dados. O plano original determinava a descoberta de como fazer os clientes passarem mais tempo e visualizarem mais páginas no site por visita.

Como se vê, manter um cliente feliz não é simplesmente uma questão de mantê-lo on-line. As distribuições de frequência do tempo gasto por sessão, divididas em a sessão ter resultado ou não em uma compra, descobriram que os tempos de sessão para a maioria de sessões que resultam em compras recaem entre os tempos de sessão de dois grupos de sessões sem compras.

Agora que isso é sabido, o problema é descobrir se esses clientes que passam muito tempo no site sem comprar nada estão apenas olhando sem compromisso ou simplesmente não conseguem encontrar o que estão procurado. A etapa seguinte é descobrir como entregar aquilo que eles estão procurando para encorajar as compras.

Realizando uma revisão do projeto final

Essa é a etapa final da metodologia do CRISP-DM e ela lhe oferece uma chance de formular suas impressões finais e intercalar as lições aprendidas durante o processo de mineração de dados.

Lista de tarefas

Você deve realizar uma breve entrevista com aqueles envolvidos de forma significativa no processo de mineração de dados. As perguntas a serem consideradas durante essas entrevistas incluem as seguintes:

- Quais são suas impressões gerais sobre o projeto?
- O que você aprendeu durante o processo, tanto em termos da mineração de dados em geral e os dados disponíveis?
- Quais partes do projeto foram bem? Onde surgiram as dificuldades? Havia informações que poderiam ter ajudado a esclarecer a confusão?

Depois de os resultados da mineração de dados terem sido implementados, você também poderá entrevistar aqueles afetados pelos resultados, como clientes ou parceiros de negócios. Sua meta aqui deve ser determinar se o projeto foi válido e se ofereceu os benefícios delimitados para serem criados.

Os resultados dessas entrevistas podem ser resumidos juntamente com suas próprias impressões do projeto em um relatório final que deve se concentrar nas lições aprendidas a partir da experiência de minerar seus armazenamentos de dados.

Exemplo de varejo eletrônico--Revisão final

Um cenário de mineração na web usando o CRISP-DM

Entrevistas com membros do projeto. O varejista eletrônico descobre que os membros do projeto mais estreitamente associados ao estudo, do início ao fim, estão, em sua maioria, entusiasmados com os resultados e aguardam projetos futuros. O grupo do banco de dados parece cuidadosamente otimista; embora apreciem a utilidade do estudo, eles apontam para a carga acrescentada aos recursos do banco de dados. Um consultor estava disponível durante o estudo, mas com o passar do tempo, será necessário outro funcionário dedicado à manutenção do banco de dados será necessário, à medida que o escopo do projeto se expande.

Entrevistas com clientes. O feedback do cliente foi amplamente positivo até o momento. Um problema que não foi bem considerado foi o impacto da mudança do design do site nos clientes estabelecidos. Depois de alguns anos, os clientes registrados desenvolveram algumas expectativas sobre como o site está organizado. O feedback de usuários registrados não é tão positivo quanto aquele de clientes não registrados e alguns realmente não gostam das mudanças. O varejista eletrônico deve estar atento a esse problema e considerar cuidadosamente se uma mudança trará novos clientes suficientes para se arriscar a perder os existentes.

Avisos

Estas informações foram desenvolvidas para os produtos e serviços oferecidos nos EUA. Este material pode estar disponível pela IBM em outros idioma. No entanto, pode ser necessário possuir uma cópia do produto ou da versão do produto no mesmo idioma para acessá-lo.

É possível que a IBM não ofereça os produtos, serviços ou recursos discutidos nesta publicação em outros países. Consulte seu representante IBM local para obter informações sobre os 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 produto, programa ou serviço funcionalmente equivalente que não infrinja nenhum direito de propriedade intelectual da IBM pode ser usado em substituição. 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 Corporation
Botafogo
Rio de Janeiro, RJ
Brasil

Para pedidos de licença relacionados a informações de Conjunto de Caracteres de Byte Duplo (DBCS), 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
IBM World Trade Asia Corporation Licensing
2-31 Roppongi 3-chome
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO "NO ESTADO EM QUE SE ENCONTRA", SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS NÃO SE LIMITANDO ÀS GARANTIAS IMPLÍCITAS DE MERCADO OU DE 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.

Referências nestas informações a Web sites não IBM são fornecidas apenas por conveniência e não representam de forma alguma um endosso a 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 utilizar ou distribuir as informações fornecidas da forma que julgar apropriada sem incorrer em qualquer obrigação para com o 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 Corporation

Botafogo
Rio de Janeiro, RJ
Brasil

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 exemplos de clientes e dados de desempenho citados são apresentados com propósitos meramente ilustrativos. Os resultados reais de desempenho podem variar, dependendo das configurações e condições operacionais específicas.

As informações relativas a produtos não IBM foram obtidas junto aos fornecedores dos respectivos produtos, de seus anúncios publicados ou de outras fontes disponíveis publicamente. A IBM não testou estes produtos e não pode confirmar a precisão de seu desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Dúvidas sobre os recursos de produtos não IBM devem ser encaminhadas diretamente a seus fornecedores.

As declarações relacionadas aos objetivos e intenções futuras da IBM estão sujeitas a alterações ou cancelamento sem aviso prévio e representam apenas metas e 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 podem incluir nomes de indivíduos, empresas, marcas e produtos. Todos estes nomes são fictícios e qualquer semelhança com nomes e endereços utilizados por uma empresa real é mera 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 empresas, produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas. Uma lista atual de marcas registradas 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 ou marcas registradas do 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 do Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou 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 do The Open Group nos Estados Unidos e/ou em outros países.

Java e todas as marcas comerciais e logotipos baseados em Java são marcas comerciais ou marcas registradas da Oracle e/ou de suas afiliadas.

Termos e condições da documentação do produto

As permissões para a utilização destas publicações são concedidas sujeitas aos termos e condições a seguir.

Aplicação

Estes termos e condições estão em adição a quaisquer termos de uso para o website IBM.

Uso pessoal

É possível reproduzir estas publicações para seu uso pessoal não comercial, desde que todos os avisos do proprietário sejam preservados. O Cliente não pode distribuir, exibir ou fazer trabalho derivado destas publicações, ou de qualquer parte delas, sem o consentimento expresso da IBM.

Uso Comercial

O Cliente pode reproduzir, distribuir e exibir estas publicações unicamente dentro de sua empresa, contanto que todos os avisos do proprietário sejam preservados. O Cliente não pode fazer trabalhos derivados destas publicações, ou reproduzir, distribuir ou exibir estas publicações ou qualquer parte delas fora da empresa, sem o consentimento expresso da IBM.

Direitas

Exceto quando expressamente concedido nesta permissão, nenhuma outra permissão, licença ou direito é concedido, seja de maneira expressa ou implícita, para as publicações ou quaisquer informações, dados, software ou outras propriedades intelectuais aqui contidas.

A IBM reserva-se o direito de retirar as permissões concedidas aqui sempre que, a seu critério, o uso das publicações seja prejudicial a seus interesses ou, conforme determinado pela IBM, as instruções acima não estejam sendo seguidas corretamente.

O Cliente não pode fazer download, exportar ou re-exportar estas informações, exceto se estiver em conformidade total com todas as leis e regulamentos aplicáveis, incluindo todas as leis e regulamentos de exportação dos Estados Unidos.

A IBM NÃO FAZ QUALQUER TIPO DE GARANTIA QUANTO AO CONTEÚDO DESTAS 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.

Índice remissivo

S

T

V