IBM SPSS Data Preparation 26

Comunicado

Antes de usar estas informações e o produto suportado por elas, leia as informações nos "Avisos" na página 7.

Informações sobre o produto

Esta edição aplica-se à versão 26, liberação 0, modificação 0 do IBM® SPSS Statistics e a todas as liberações e modificações subsequentes até que seja indicado de outra forma em novas edições.

Índice

Preparação de dados

Os seguintes recursos de preparação de dados estão incluídos naSPSS Statistics Professional Edition ou na opção de preparação de dados.

Introdução à preparação de dados

À medida que aumenta a força dos sistemas de computação, os interesses por informações crescem proporcionalmente, conduzindo a mais e mais coletas de dados — mais casos, mais variáveis e mais erros de entrada de dados. Esses erros são o flagelo das previsões de modelo preditivo que são o objetivo final de data warehousing, portanto, é necessário manter os dados "limpos." No entanto, a quantidade de dados armazenados cresceu além da capacidade de verificar os casos manualmente, o que é vital para implementar processos automatizados para validar dados.

Os do módulo complementar de preparação de dados permitem que você identifique casos incomuns, casos inválidos, variáveis e valores de dados em seu conjunto de dados ativo e prepare dados para a modelagem.

Uso de procedimentos de preparação de dados

Seu uso de procedimentos de preparação de dados depende de suas necessidades específicas. Uma rota típica, após carregar seus dados, é:

Preparação de metadados
: Revise as variáveis em seu arquivo de dados e determine seus valores, rótulos e níveis de medição válidos. Identifique combinações de valores da variável que são impossíveis, mas comumente codificados incorretamente. Defina regras de validação com base nestas informações. Essa pode ser uma tarefa demorada, mas vale o esforço, se você precisar validar arquivos de dados com atributos semelhantes regularmente.

Validação de dados
: Execute verificações básicas e verificações em regras de validação definidas para identificar casos, variáveis e valores de dados inválidos. Quando forem encontrados dados inválidos, investigue e corrija a causa. Isso pode requerer outro passo por meio da preparação de metadados.

Preparação de modelo
: Use a preparação de dados automatizada para obter transformações dos campos originais que irão melhorar a construção de modelo. Identifique possíveis valores discrepantes estatísticos que podem causar problemas para muitos modelos preditivos. Alguns valores discrepantes são o resultado de valores da variável inválidos que não foram identificados. Isso pode requerer outro passo por meio da preparação de metadados.

Quando seu arquivo de dados estiver "limpo", você estará pronto para criar modelos de outros módulos complementares.

Identificar casos incomuns

O procedimento de detecção de anomalias procura casos incomuns com base em desvios das normas de seus grupos de clusters. O procedimento foi projetado para detectar rapidamente casos incomuns para propósitos de auditoria de dados no passo de análise de dados exploratória, antes de qualquer análise de dados inferencial. Esse algoritmo foi projetado para detecção de anomalias genéricas; ou seja, a definição de um caso anômalo não é específica de nenhum aplicativo específico, como a detecção de padrões de pagamento incomuns no segmento de mercado de assistência médica ou detecção de lavagem de dinheiro no segmento de mercado de finanças, no qual a definição de uma anomalia pode ser bem definida.

Exemplo:

Um analista de dados contratado para construir modelos preditivos para resultados de tratamento de AVC está preocupado com a qualidade de dados, porque esses modelos podem ser sensíveis a observações incomuns. Algumas dessas observações distantes representam casos realmente exclusivos e, portanto, não são apropriadas para predição, enquanto outras observações são causadas por erros de entrada de dados nos quais os valores estão tecnicamente "corretos" e, portanto, não podem ser capturados por procedimentos de validação de dados. O procedimento Identificar casos incomuns localiza e relata esses valores discrepantes para que o analista possa decidir como tratá-los.

Estatísticas

O procedimento produz grupos de peers, normas do grupo de peers para variáveis contínuas e categóricas, índices de anomalia com base em desvios de normas do grupo de peers, e valores de impacto de variável para variáveis que mais contribuem com um caso que está sendo considerado incomum.

Considerações de Dados

Dados. Este procedimento funciona com variáveis contínuas e categóricas. Cada linha representa uma observação distinta e cada coluna representa uma variável distinta na qual os grupos de peers são baseados. Uma variável de identificação de caso pode estar disponível no arquivo de dados para marcar a saída, mas não será usada na análise. Os valores omissos são permitidos. A variável de ponderação, se especificada, é ignorada.

O modelo de detecção pode ser aplicado a um novo arquivo de dados de teste. Os elementos dos dados de teste devem ser iguais aos elementos dos dados de treinamento. E, dependendo das configurações do algoritmo, o tratamento de valor omisso que é usado para criar o modelo pode ser aplicado ao arquivo de dados de teste antes da escoragem.

Ordem de casos. Observe que a solução pode depender da ordem dos casos. Para minimizar os efeitos da ordem, ordene aleatoriamente os casos. Para verificar a estabilidade de uma determinada solução, talvez você queira obter várias soluções diferentes com casos ordenados em diferentes ordens aleatórias. Em situações com tamanhos de arquivos extremamente grandes, podem ser feitas várias execuções com uma amostra de casos ordenados em diferentes ordens aleatórias.

Suposições. O algoritmo considera que todas as variáveis são inconstantes e independentes e que nenhum caso possui valores omissos para qualquer uma das variáveis de entrada. Cada variável contínua é considerada como tendo uma distribuição normal (Gaussiana) e cada variável categórica é considerada como tendo uma distribuição multinomial. O teste interno empírico indica que o procedimento é bastante robusto a violações da suposição de independência e das suposições distributivas, mas esteja ciente de como essas suposições são atendidas.

Identificando casos incomuns

1. Nos menus, escolha:
   Dados > **Identificar casos incomuns...**
2. Selecione pelo menos uma variável de análise.
3. Opcionalmente, escolha uma variável identificadora de caso para ser usada na identificação de saída.
4. Dê um clique em **Aplicar**.

Campos com nível de medição desconhecido

O alerta do nível de medição é exibido quando o nível de medição de uma ou mais variáveis (campos) no conjunto de dados é desconhecido. Como o nível de medição afeta o cálculo de resultados para este procedimento, todas as variáveis devem ter um nível de medição definido.

Verificar dados
Lê os dados no conjunto de dados ativo e designa o nível de medição padrão para quaisquer campos com um nível de medição desconhecido atualmente. Se o conjunto de dados for grande, isso poderá demorar algum tempo.

Designe manualmente
Lista todos os campos com um nível de medição desconhecido. É possível designar um nível de medição a esses campos. Também é possível designar um nível de medição no painel Lista de Variáveis do Editor de Dados.

Como o nível de medição é importante para esse procedimento, não é possível executar o procedimento até que todos os campos tenham um nível de medição definido.

Identifique casos incomuns: saída

A caixa de diálogo Saída fornece opções para a geração do resultado tabular.

Lista de casos incomuns e motivos pelos quais eles são considerados incomuns
Quando selecionada, essa opção produz três tabelas:

- A lista de índice de casos de anomalia exibe casos que são identificados como incomuns e exibe seus valores de índice de anomalia correspondentes.
- A lista de IDs de peers de casos de anomalia exibe casos incomuns e informações referentes a seus grupos de peers correspondentes.
- A lista de razões de anomalia exibe o número do caso, a variável de razão, o valor de impacto da variável, o valor da variável e a norma da variável para cada razão.

Todas as tabelas são ordenadas por índice de anomalia em ordem decrescente. Além disso, os IDs dos casos são exibidos se a variável identificadora de caso for especificada na caixa de diálogo Variáveis.

Resumos
Os controles nesse grupo produzem sumarizações de distribuição.

Normas de grupo de peers
Essa opção exibe a tabela de normas de variáveis contínuas (se alguma variável contínua for usada na análise) e a tabela de normas de variáveis categóricas (se alguma variável categórica for usada na análise). A tabela de normas de variáveis contínuas exibe a média e o desvio padrão de cada variável contínua para cada grupo de peers. A tabela de normas de variável categórica exibe o modo (categoria mais popular), frequência e porcentagem de frequência de cada variável categórica para cada grupo de peers. A média de uma variável contínua e o modo de uma variável categórica são usados como os valores de norma na análise.

Índices de anomalia
A sumarização de índice de anomalia exibe estatísticas descritivas para o índice de anomalia dos casos que são identificados como os mais incomuns.

Ocorrência de motivo por variável de análise
Para cada razão, a tabela exibe a frequência e a porcentagem de frequência de ocorrência de cada variável como uma razão. A tabela também relata as estatísticas descritivas do impacto de cada variável. Se o número máximo de razões estiver configurado como 0 na guia Opções, essa opção não estará disponível.

Casos processados
A sumarização de processamento de caso exibe as contagens e porcentagens de contagens para todos os casos no conjunto de dados ativo, os casos incluídos e excluídos na análise e os casos em cada grupo de peers.

Identifique casos incomuns: salvar

A caixa de diálogo Salvar fornece opções de salvamento de modelo e de variável.

Salvar variáveis

Os controles nesse grupo permitem salvar variáveis de modelo no conjunto de dados ativo. Também é possível optar por substituir variáveis existentes cujos nomes entram em conflito com as variáveis a serem salvas.

Índice de anomalia

Salva o valor do índice de anomalia para cada caso para uma variável com o nome especificado.

Grupos de peers

Salva o ID do grupo de peers, a contagem de casos e o tamanho como uma porcentagem para cada caso para variáveis com o nome raiz especificado. Por exemplo, se o nome raiz *Peer* for especificado, as variáveis *Peerid*, *PeerSize* e *PeerPctSize* serão geradas. *Peerid* é o ID do grupo de peers do caso, *PeerSize* é o tamanho do grupo e *PeerPctSize* é o tamanho do grupo como uma porcentagem.

Razões

Salva conjuntos de variáveis de razão com o nome raiz especificado. Um conjunto de variáveis de razão consiste no nome da variável como a razão, sua medida de impacto da variável, seu próprio valor e o valor da norma. O número de conjuntos depende do número de razões solicitadas na guia Opções. Por exemplo, se o nome raiz *Reason* for especificado, as variáveis *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k* e *ReasonNorm_k* serão geradas, em que *k* é a razão *k*. Essa opção não estará disponível se o número de razões estiver configurado como 0.

Substituir variáveis existentes que têm o mesmo nome ou nome raiz

Quando selecionadas, as variáveis existentes cujos nomes entram em conflito com as variáveis a serem salvas são substituídas.

Exportar arquivo de modelo

Permite salvar o modelo em um arquivo XML externo.

Identifique casos incomuns: valores omissos

A caixa de diálogo Valores Omissos é usada para controlar a manipulação de valores omissos do sistema e do usuário.

Excluir valores omissos da análise

Casos com valores omissos são excluídos da análise.

Incluir valores omissos na análise

Os valores omissos de variáveis contínuas são substituídos por suas médias globais correspondentes e as categorias omissas de variáveis categóricas são agrupadas e tratadas como uma categoria válida. As variáveis processadas são então usadas na análise. Opcionalmente, é possível solicitar a criação de uma variável adicional que representa a proporção de variáveis omissas em cada caso e usar essa variável na análise.

Identifique casos incomuns: opções

A caixa de diálogo Opções inclui as configurações para critérios de caso incomuns e define um intervalo para o número de grupos de peers.

Critérios para identificar casos incomuns

Essas configurações a seguir determinam quantos casos são incluídos na lista de anomalias.

Porcentagem de casos com os maiores valores de índice de anomalia

Especifique um número positivo que seja menor ou igual a 100.

Número fixo de casos com os maiores valores de índice de anomalia
Especifique um número inteiro positivo que seja menor ou igual ao número total de casos no conjunto de dados ativo que são usados na análise.

Identificar apenas casos cujo valor de índice de anomalia atende ou excede um valor mínimo
Especifique um número não negativo. Um caso é considerado anômalo se seu valor de índice de anomalia for maior ou igual ao ponto de corte especificado. Essa opção é usada junto com as opções **Porcentagem de casos** e **Número fixo de casos**. Por exemplo, se você especificar um número fixo de 50 casos e um valor de corte de 2, a lista de anomalia consistirá, no máximo, de 50 casos, cada um com um valor de índice de anomalia maior ou igual a 2.

Número de grupos de peers
O procedimento procura o melhor número de grupos de peers entre os valores mínimo e máximo especificados. Os valores devem ser números inteiros positivos e o mínimo não deve exceder o máximo. Quando os valores especificados forem iguais, o procedimento considerará um número fixo de grupos de peers.

Nota: Dependendo da quantidade de variação em seus dados, pode haver situações nas quais o número de grupos de pares que os dados podem suportar seja menor que o número especificado como o mínimo. Nessa situação, o procedimento pode produzir um número menor de grupos de peers.

Número máximo de motivos
Uma razão consiste na medida de impacto da variável, no nome da variável para essa razão, o valor da variável e no valor do grupo de peers correspondente. Especifique um número inteiro não negativo; se esse valor for igual ou exceder o número de variáveis processadas que são usadas na análise, todas as variáveis serão mostradas.

Recursos adicionais do comando DETECTANOMALY

O idioma da sintaxe de comando também permite:

- Omita algumas variáveis no conjunto de dados ativo da análise sem especificar explicitamente todas as variáveis de análise (usando o subcomando `EXCEPT`).
- Especifique um ajustamento para balancear a influência de variáveis contínuas e categórica (usando a palavra-chave `MLWEIGHT` no subcomando `CRITERIA`).

Consulte a *Referência da sintaxe de comando* para obter informações de sintaxe completa.

Avisos

Essas informações foram desenvolvidas para produtos e serviços oferecidos nos Estados Unidos. Esse material pode estar disponível a partir da IBM em outros idiomas. No entanto, pode ser necessário possuir uma cópia do produto ou da versão do produto nesse idioma para acessá-lo.

É possível que a IBM não ofereça produtos, serviços ou recursos discutidos neste documento em outros países. Consulte um representante IBM local para obter informações sobre produtos e serviços disponíveis atualmente em sua área. Qualquer referência a produtos, programas ou serviços IBM não significa que apenas produtos, programas ou serviços IBM possam ser utilizados. Qualquer produto, programa ou serviço funcionalmente equivalente, que não infrinja nenhum direito de propriedade intelectual da IBM poderá ser utilizado em substituição a este produto, programa ou serviço. 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 podem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
CEP 22290-240
Rio de Janeiro, RJ
Brasil

Para pedidos de licença relacionados a informações de DBCS (Conjunto de Caracteres de Byte Duplo), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie pedidos de licença, por escrito, para:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

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 NÃO-VIOLAÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias explícitas ou implícitas em certas transações; portanto, esta instrução 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 o(s) produto(s) e/ou programa(s) descritos nesta publicação, sem aviso prévio.

Qualquer referência nestas informações a websites 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 para esse produto IBM e o uso desses websites é de inteira responsabilidade do Cliente.

A IBM por usar 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 o mesmo com o objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) o uso mútuo de informações trocadas, devem entrar em contato com:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
CEP 22290-240
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 dados de desempenho e do Cliente citados são apresentados apenas para propósitos ilustrativos. Resultados de desempenho reais podem variar dependendo das configurações específicas e das condições operacionais.

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 esses produtos e não pode confirmar a precisão de desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Perguntas sobre os recursos de produtos não IBM devem ser endereçadas aos fornecedores desses produtos.

Instruções relativas à direção futura ou intento da IBM estão sujeitas a mudança ou retirada sem aviso e representam metas e objetivos apenas.

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 assuntos, empresas, marcas e produtos. Todos esses nomes são fictícios e qualquer semelhança com pessoas ou empresas reais é mera coincidência.

LICENÇA DE COPYRIGHT:

Estas informações contêm programas de aplicativos de amostra na linguagem fonte, ilustrando as técnicas de programação em diversas plataformas operacionais. O Cliente pode copiar, modificar e distribuir estes programas de amostra sem a necessidade de pagar à IBM, com objetivos de desenvolvimento, utilização, marketing ou distribuição de programas aplicativos em conformidade com a interface de programação de aplicativo para a plataforma operacional para a qual os programas de amostra são criados. Esses exemplos não foram testados completamente em todas as condições. Portanto, a IBM não pode garantir ou implicar a confiabilidade, manutenção ou função destes programas. Os programas de amostra são fornecidos "NO ESTADO EM QUE SE ENCONTRAM", sem garantia de qualquer tipo. A IBM não será responsabilizada por quaisquer danos decorrentes do uso dos programas de amostra.

Cada cópia ou parte destes programas de amostra ou qualquer trabalho derivado deve incluir um aviso de copyright com os dizeres:

Marcas comerciais

IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em muitos países no mundo todo. Outros nomes de produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas. ma lista atual de marcas comerciais da IBM está disponível na web em "Copyright and trademark information" em www.ibm.com/legal/copytrade.shtml.

Adobe, o logotipo Adobe, PostScript e o logotipo PostScript são marcas registradas ou marcas comerciais da Adobe Systems Incorporated nos Estados Unidos e/ou em outros países.

Intel, o logotipo Intel, Intel Inside, o logotipo Intel Inside, Intel Centrino, o logotipo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou de suas subsidiárias nos Estados Unidos e em outros países.

Linux é uma marca registrada da Linus Torvalds nos Estados Unidos, e/ou em outros países.

Microsoft, Windows, Windows NT e o logotipo Windows são marcas comerciais da Microsoft Corporation nos Estados Unidos e/ou em outros países.

UNIX é uma marca registrada da The Open Group nos Estados Unidos e em outros países.

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

Índice Remissivo

G

I

M

V

Impresso no Brasil