IBM SPSS Data Preparation 29

Poznámka

Před použitím těchto informací a produktu, který podporují, si přečtěte informace v tématu “Upozornění” na stránce 25.

Informace o produktu

Toto vydání se vztahuje k verzi 29, vydání 0, modifikaci 1 produktu IBM® SPSS Statistics a ke všem následujícím vydáním a modifikacím, dokud nebude v nových vydáních uvedeno jinak.

Obsah

Kapitola 1. Příprava dat..... 1
Úvod do přípravy dat..... 1
Použití procedur přípravy dat..... 1
Pravidla ověření platnosti..... 1
Načíst předdefinovaná pravidla ověření platnosti..... 1
Definovat pravidla ověření platnosti..... 2
Ověřit data..... 3
Ověřit základní kontroly dat..... 4
Ověřit pravidla pro jednoměnná data..... 5
Ověřit pravidla pro křížovou proměnnou dat..... 5
Ověřit výstup dat..... 5
Ověřit uložení dat..... 6
Automatizovaná příprava dat..... 6
Chcete-li získat automatické přípravy dat..... 7
Chcete-li získat interaktivní přípravu dat..... 7
Karta Pole..... 8
Karta Nastavení..... 8
Karta Analýza..... 12
Zpětná hodnocení transformace..... 18
Identifikace neobvyklých případů..... 18
Identifikace neobvyklých výstupních případů..... 19
Identifikovat případy, kdy nejsou běžné případy..... 20
Identifikace neobvyklých případů Chybějící hodnoty..... 20
Identifikace neobvyklých voleb případů..... 20
Další funkce příkazu DETECTANOMALY..... 21
Optimální ukotvení..... 21
Optimálně vypalovací výstup..... 21
Optimální ukládání do neaktivního stavu..... 22
Optimální vychycení chybějících hodnot..... 22
Optimální volby ukotvení..... 22
Další funkce příkazu OPTIMAL BINNING..... 23

Upozornění..... 25
Ochranné známky..... 26

Rejstřík..... 29

Kapitola 1. Příprava dat

Produkt Base Edition obsahuje následující funkce pro přípravu dat.

Úvod do přípravy dat

Při zvyšování výkonu výpočetních systémů se zvyšuje počet žádostí o informace, což vede k většímu počtu datových kolekcí-více případů, více proměnných a více chyb při zadávání dat. Tyto chyby jsou bane prediktivních modelových prognóz, které jsou konečným cílem ukládání do datového skladu, takže potřebujete uchovat data "čistá". Množství skladovaných dat však dosud nevzrostlo, než je schopnost ověřovat případy ručně, že je nezbytně nutné implementovat automatizované procesy pro ověřování dat.

Příprava dat vám umožňuje identifikovat neobvyklé případy a neplatné případy, proměnné a datové hodnoty ve vaší aktivní datové sadě a připravit data pro modelování.

Použití procedur přípravy dat

Použití procedur pro přípravu dat závisí na vašich konkrétních potřebách. Typická trasa po načtení vašich dat je:

- **Příprava metadat.** Zkontrolujte proměnné ve svém datovém souboru a určete jejich platné hodnoty, štítky a úrovně měření. Identifikovat kombinace hodnot proměnných, které jsou neproveditelné, ale často chybně kódované. Definujte ověřovací pravidla na základě těchto informací. Může se jednat o časově náročnou úlohu, ale za vynaloženém úsilí se vyplatí, pokud potřebujete pravidelně ověřovat datové soubory s podobnými atributy.
- **Ověření dat.** Spusťte základní kontroly a zkontrolujte definovaná pravidla ověření, abyste identifikovali neplatné případy, proměnné a datové hodnoty. Jsou-li nalezena neplatná data, prozkoumejte a opravte příčinu. To může vyžadovat další krok prostřednictvím přípravy metadat.
- **Příprava modelu.** Využijte automatizované zpracování dat k získání transformací původních polí, která zlepší sestavování modelu. Identifikujte potenciální statistické odlehlé hodnoty, které mohou způsobit problémy pro mnoho prediktivních modelů. Některé odlehlé hodnoty jsou výsledkem neplatných hodnot proměnných, které nebyly identifikovány. To může vyžadovat další krok prostřednictvím přípravy metadat.

Jakmile je datový soubor "čistý", jste připraveni sestavit modely z jiných doplňkových modulů.

Pravidla ověření platnosti

Pravidlo se používá k určení, zda je případ platný. Existují dva typy pravidel pro ověření platnosti:

- **Jednotlivá proměnná pravidla.** Pravidla jednotlivých proměnných se skládají z pevné sady kontrol, které se vztahují k jedné proměnné, jako jsou například kontroly hodnot mimo rozsah hodnot. Pro pravidla s jednou proměnnou mohou být platné hodnoty vyjádřeny jako rozsah hodnot nebo seznam přijatelných hodnot.
- **Meziproměnlivá pravidla.** Pravidla proměnných mezi proměnnými jsou uživatelem definovaná pravidla, která lze aplikovat na jedinou proměnnou nebo kombinaci proměnných. Pravidla proměnných mezi proměnnými jsou definována logickým výrazem, který označuje neplatné hodnoty.

Pravidla ověření platnosti jsou uložena do datového slovníku vašeho datového souboru. To vám umožní zadat pravidlo jednou a pak je znovu použít.

Načíst předdefinovaná pravidla ověření platnosti

Můžete rychle získat sadu pravidel pro ověření platnosti po použití načtením předdefinovaných pravidel z externího datového souboru zahrnutého v instalaci.

Načíst předdefinovaná pravidla ověření platnosti

1. Z nabídky vyberte:

   Data > **Ověření** > **Načíst předdefinovaná pravidla ...**

Alternativně můžete použít Průvodce kopírováním vlastností dat k načtení pravidel ze všech datových souborů.

Definovat pravidla ověření platnosti

Dialogové okno Definovat pravidla pro ověření vám umožňuje vytvořit a zobrazit pravidla pro ověření jednotlivé proměnné a křížových proměnných.

Chcete-li vytvořit a zobrazit pravidla ověření platnosti

1. Z nabídky vyberte:

   Data > **Ověření** > **Definovat pravidla ...**

   Dialogové okno je naplněno pravidly pro ověření jednotlivých proměnných a křížových proměnných, která se čtou ze slovníku dat. Když nejsou žádná pravidla, nové pravidlo zástupného symbolu, které můžete upravit, aby vyhovovalo vašemu účelu, se vytvoří automaticky.
2. Vyberte jednotlivá pravidla na kartách Pravidla jednotlivých proměnných a Pravidla pro křížovou proměnnou, chcete-li zobrazit a upravit jejich vlastnosti.

Definovat pravidla s jednou proměnnou

Ouško Single-Variable Rules vám umožňuje vytvářet, zobrazovat a upravovat pravidla pro ověření platnosti jedné proměnné.

Pravidla. Seznam zobrazuje pravidla ověření platnosti jednotlivých proměnných podle názvu a typu proměnné, na kterou lze pravidlo použít. Když je dialogové okno otevřeno, zobrazí pravidla definovaná v datovém slovníku nebo, nejsou-li momentálně definována žádná pravidla, pravidlo zástupného symbolu nazvané "Pravidlo jedné proměnné 1." Níže uvedená tlačítka se zobrazí pod seznamem pravidel:

- **Nový.** Přidá nový záznam do dolní části seznamu pravidel. Pravidlo je vybráno a přiřazen název "SingleVarRule *n*," kde *n* je celé číslo, takže název nového pravidla je jedinečný v rámci jedné proměnné a pravidel napříč proměnnými.
- **Duplikovat.** Přidá kopii vybraného pravidla do dolní části seznamu pravidel. Název pravidla se upraví tak, aby byl jedinečný v rámci jednoproměnných a pravidel proměnných mezi proměnnými. Například, pokud duplikujete "SingleVarRule 1," název prvního duplicitního pravidla bude "Copy of SingleVarRule 1," the second would be "Copy (2) of SingleVarRule 1," and so on.
- **Odstranit.** Odstraní vybrané pravidlo.

Definice pravidla. Tyto ovládací prvky umožňují zobrazit a nastavit vlastnosti pro vybrané pravidlo.

- **Název.** Název pravidla musí být jedinečný v rámci jedné proměnné a pravidel pro více proměnných.
- **Typ.** Jedná se o typ proměnné, na kterou lze pravidlo použít. Vyberte z voleb **Čísla**, **Řetězec**a **Datum**.
- **Formát.** To vám umožňuje vybrat formát data pro pravidla, která lze použít na proměnné data.
- **Platné hodnoty.** Platné hodnoty můžete zadat buď jako rozsah, nebo jako seznam hodnot.

  Definice rozsahu

  Ovládací prvky definice rozsahu umožňují určit platný rozsah. Hodnoty mimo rozsah jsou označeny jako neplatné.

  Chcete-li určit rozsah, zadejte minimální nebo maximální hodnoty, nebo obojí. Ovládací prvky zaškrtávacího políčka umožňují označit neoznačené a necelé hodnoty v rámci rozsahu.

  Definice seznamu

  Ovládací prvky definice seznamu umožňují definovat seznam platných hodnot. Hodnoty, které nejsou zahrnuty v seznamu, jsou označeny jako neplatné.

Zadejte hodnoty seznamu do mřížky. Zaškrtávací políčko určuje, zda mají být při kontrole hodnot řetězcových dat v seznamu přijatelných hodnot rozlišována velká a malá písmena.

- **Povolit uživateli-chybějící hodnoty.** Určuje, zda jsou uživatelé-chybějící hodnoty označeny příznakem jako neplatné.
- **Povolit systémem chybějící hodnoty.** Určuje, zda jsou systémové hodnoty, které chybí, označeny jako neplatné. To neplatí pro typy řetězcových pravidel.
- **Povolit prázdné hodnoty.** Určuje, zda jsou hodnoty řetězce prázdné (to znamená úplné prázdné) označeny jako neplatné. To neplatí pro neřetězcové typy pravidel.

Definovat pravidla pro více proměnných

Ouško Cross-Variable Rules vám umožňuje vytvořit, zobrazit a upravit pravidla pro ověření křížových proměnných.

Pravidla. Seznam zobrazuje pravidla ověření pro více proměnných podle názvu. Když je dialogové okno otevřeno, zobrazí se pravidlo zástupného symbolu "CrossVarRule 1." Níže uvedená tlačítka se zobrazí pod seznamem pravidel:

- **Nový.** Přidá nový záznam do dolní části seznamu pravidel. Pravidlo je vybráno a přiřazen název "CrossVarRule *n*," kde *n* je celé číslo, takže název nového pravidla je jedinečný v rámci jedné proměnné a pravidel napříč proměnnými.
- **Duplikovat.** Přidá kopii vybraného pravidla do dolní části seznamu pravidel. Název pravidla se upraví tak, aby byl jedinečný v rámci jednoproměnných a pravidel proměnných mezi proměnnými. Například, pokud duplikujete "CrossVarRule 1", název prvního duplicitního pravidla bude "Kopie z CrossVarRule 1," druhá by byla "Kopie (2) z CrossVarRule 1," atd.
- **Odstranit.** Odstraní vybrané pravidlo.

Definice pravidla. Tyto ovládací prvky umožňují zobrazit a nastavit vlastnosti pro vybrané pravidlo.

- **Název.** Název pravidla musí být jedinečný v rámci jedné proměnné a pravidel pro více proměnných.
- **Logický výraz.** To je v podstatě definice pravidla. Měli byste kódovat výraz tak, aby se neplatné případy vyhodnotily na hodnotu 1.

Sestavení výrazů

1. Chcete-li sestavit výraz, vložte komponenty do pole Výraz nebo zadejte přímo do pole Výraz.

- Funkce nebo běžně používané systémové proměnné můžete vložit výběrem skupiny ze seznamu skupin funkcí a dvojitým klepnutím na funkci nebo proměnnou v seznamu Funkce a speciální proměnné (nebo vyberte funkci či proměnnou a klepněte na tlačítko **Vložit**). Zadejte hodnoty pro všechny parametry označené otazníky (platí pouze pro funkce). Skupina funkcí označená **Vše** poskytuje seznam všech dostupných funkcí a systémových proměnných. Krátký popis aktuálně vybrané funkce nebo proměnné se zobrazí ve vyhrazené oblasti v dialogovém okně.
- Řetězcové konstanty musí být uzavřeny v uvozovkách nebo apostrofech.
- Pokud hodnoty obsahují desetinná čísla, musí být jako desetinný indikátor použita tečka (.).

Ověřit data

Dialogové okno Ověřit data vám umožňuje identifikovat podezřelé a neplatné případy, proměnné a datové hodnoty v aktivní datové sadě.

Příklad. Analytik dat musí poskytnout svému klientovi měsíční sestavu spokojenosti zákazníků. Data, která obdrží každý měsíc, musí být zkontrolována kvalitou pro nekompletní ID zákazníků, hodnoty proměnných, které jsou mimo rozsah, a kombinace hodnot proměnných, které jsou obvykle zadány v chybě. Dialogové okno Ověřit data umožňuje analytikovi určit proměnné, které jedinečně identifikují zákazníky, definují pravidla jediné proměnné pro platné rozsahy proměnných a definují pravidla mezi proměnnými, aby bylo možné zachytit nemožné kombinace. Procedura vrátí zprávu o problémových případech a proměnných. Kromě toho mají data každý měsíc stejné datové prvky, takže analytik je schopen aplikovat pravidla na nový datový soubor příští měsíc.

Statistika. Procedura vytvoří seznamy proměnných, případů a hodnot dat, které selžou při různých kontrolách, počtů porušení pravidel jedné proměnné a proměnných napříč proměnnými a jednoduché deskriptivní souhrny proměnných analýzy.

Váhy. Procedura ignoruje specifikaci proměnné váhy a místo toho s ní zachází jako s jakoukoli jinou analýzou.

Ověřit data

1. Z nabídky vyberte:

   Data > **Ověření** > **Ověřit data ...**
2. Vyberte jednu nebo více proměnných analýzy pro ověření pomocí základních kontrol proměnných nebo pomocí pravidel pro ověření platnosti jedné proměnné.

   Případně můžete:
3. Klepněte na kartu **Cross-Variable Rules** a použijte jedno nebo více pravidel napříč proměnnými.

Volitelně můžete:

- Vyberte jednu nebo více identifikačních proměnných případu, které se mají zkontrolovat u duplicitních nebo neúplných ID. Proměnné ID případu se také používají k označení výstupu na základě případu. Jsou-li zadány dvě nebo více proměnných ID případu, je jejich kombinace považována za identifikátor případu.

Pole s neznámou úrovní měření

Výstraha na úrovni měření se zobrazí, když je úroveň měření pro jednu nebo více proměnných (polí) v datové sadě neznámá. Jelikož úroveň měření ovlivňuje výpočet výsledků pro tuto proceduru, všechny proměnné musí mít definovanou úroveň měření.

Data skenování. Přečte data v aktivní datové sadě a přiřadí výchozí úroveň měření k jakýmkoli polím s momentálně neznámou úrovní měření. Je-li datová sada velká, může to nějakou dobu trvat.

Přiřadit ručně. Otevře dialogové okno se seznamem všech polí s neznámou úrovní měření. Toto dialogové okno můžete použít k přiřazení úrovně měření k těmto polím. Úroveň měření můžete také přiřadit v pohledu Proměnné v editoru dat.

Vzhledem k tomu, že úroveň měření je pro tuto proceduru důležitá, nemůžete přistupovat k dialogovému oknu pro spuštění této procedury, dokud nebude mít všechna pole definovanou úroveň měření.

Ověřit základní kontroly dat

Karta Základní kontroly vám umožňuje vybrat základní kontroly pro proměnné analýzy, identifikátory případů a celé případy.

Proměnné analýzy. Pokud jste na kartě Proměnné vybrali libovolnou analytickou proměnnou, můžete vybrat kteroukoli z následujících kontrol jejich platnosti. Zaškrtávací políčko vám umožňuje zapnout nebo vypnout kontroly.

- **Maximální procentní část chybějících hodnot.** Hlásí proměnné analýzy s procentuálním podílem chybějících hodnot, které jsou větší než uvedená hodnota. Uvedená hodnota musí být kladné číslo menší nebo rovné 100.
- **Maximální procentní část případů v jedné kategorii.** Jsou-li nějaké proměnné analýzy kategorické, tato volba hlásí proměnné kategorické analýzy s procentem případů představujících jedinou nechybějící kategorii, která je větší než uvedená hodnota. Uvedená hodnota musí být kladné číslo menší nebo rovné 100. Procentní část je založena na případech s nechybějícími hodnotami proměnné.
- **Maximální procentní část kategorií s počtem 1.** Jsou-li nějaké proměnné analýzy kategorické, tato volba uvádí proměnné kategoriální analýzy, ve kterých je procentní část kategorií proměnných obsahujících pouze jeden případ vyšší než uvedená hodnota. Uvedená hodnota musí být kladné číslo menší nebo rovné 100.
- **Minimální koeficient odchylky.** Pokud jsou některé proměnné analýzy stupnice, tato volba uvádí proměnné analýzy měřítka, ve kterých je absolutní hodnota variačního koeficientu menší než uvedená

hodnota. Tato volba se vztahuje pouze na proměnné, ve kterých je střední hodnota nenulová. Uvedená hodnota musí být nezáporné číslo. Při zadání hodnoty 0 bude provedena kontrola koeficientu změny.

- **Minimální směrodatná odchylka.** Jsou-li některé proměnné analýzy měřítka, tato volba uvádí proměnné analýzy měřítka, jejichž směrodatná odchylka je menší než uvedená hodnota. Uvedená hodnota musí být nezáporné číslo. Při zadání hodnoty 0 bude provedena kontrola směrodatné odchylky.

Identifikátory případů. Pokud jste na kartě Proměnné vybrali proměnné identifikátoru případu, můžete vybrat kteroukoli z následujících kontrol platnosti jejich platnosti.

- **Označit nekompletní ID.** Tato volba nahlásí případy s neúplnými identifikátory případu. V případě konkrétního případu je identifikátor považován za neúplný, pokud je hodnota některé proměnné ID prázdná nebo chybí.
- **Označte duplicitní ID.** Tato volba uvádí případy s duplicitními identifikátory případu. Neúplné identifikátory jsou vyloučeny ze sady možných duplicit.

Označit prázdné případy příznakem. Tato volba ohlásí případy, ve kterých jsou všechny proměnné prázdné nebo prázdné. Za účelem identifikace prázdných případů můžete zvolit použití všech proměnných v souboru (kromě proměnných ID) nebo pouze proměnných analýzy definovaných na kartě Proměnné.

Ověřit pravidla pro jednoměnná data

Na kartě Pravidla jednotlivých proměnných se zobrazují dostupná pravidla ověření jednotlivých proměnných a umožňují jejich použití na proměnné analýzy. Chcete-li definovat další pravidla s jednou proměnnou, klepněte na volbu **Definovat pravidla**. Další informace naleznete v tématu “Definovat pravidla s jednou proměnnou” na stránce 2 .

Proměnné analýzy. Seznam zobrazuje proměnné analýzy, shrnuje jejich distribuce a uvádí počet pravidel použitých pro každou proměnnou. Všimněte si, že hodnoty user-and system-missing values are not included in the summaries. Rozevírací seznam Zobrazení řídí, které proměnné jsou zobrazeny; můžete si vybrat z voleb **Všechny proměnné**, **Číselné proměnné**, **Proměnné řetězce**a **Proměnné data**.

Pravidla. Chcete-li použít pravidla pro analýzu proměnných, vyberte jednu nebo více proměnných a zaškrtněte všechna pravidla, která chcete použít v seznamu pravidel. Seznam pravidel zobrazuje pouze pravidla, která jsou vhodná pro vybrané proměnné analýzy. Jsou-li například vybrány číselné proměnné analýzy, zobrazí se pouze numerická pravidla; pokud je vybrána řetězcová proměnná, zobrazí se pouze řetězcová pravidla. Nejsou-li vybrány žádné proměnné analýzy, nebo mají smíšené datové typy, nejsou zobrazena žádná pravidla.

Distribuce proměnných. Souhrny distribuce zobrazené v seznamu Proměnné analýzy mohou být založeny na všech případech nebo na skenování prvních *n* případů, jak je uvedeno v textovém poli Případy. Klepnutí na volbu **Znovu skenovat** aktualizuje souhrny distribuce.

Ověřit pravidla pro křížovou proměnnou dat

Karta Pravidla pro více proměnných zobrazuje dostupná pravidla napříč proměnnými a umožňuje vám je použít na svá data. Chcete-li definovat další pravidla mezi proměnnými, klepněte na volbu **Definovat pravidla**. Další informace naleznete v tématu “Definovat pravidla pro více proměnných” na stránce 3 .

Ověřit výstup dat

Sestava Casewise. Pokud jste aplikovali jakákoli pravidla ověření na jednu proměnnou nebo na více proměnných, můžete požádat o sestavu, která vypíše pro jednotlivé případy narušení pravidla ověření platnosti.

- **Minimální počet narušení.** Tato volba uvádí minimální počet případů porušení pravidel požadovaných pro případ, který se má zahrnout do sestavy. Uveďte kladné celé číslo.
- **Maximální počet případů.** Tato volba uvádí maximální počet případů zahrnutých v sestavě případu. Zadejte kladné celé číslo menší nebo rovné 1000.

Pravidla pro ověření platnosti s jednou proměnnou. Pokud jste aplikovali jakákoli pravidla pro ověření platnosti jedné proměnné, můžete zvolit, jak se mají zobrazit výsledky, nebo zda se mají zobrazit vůbec.

- **Shrnout narušení podle proměnné analýzy.** Pro každou proměnnou analýzy tato volba zobrazí všechna porušení pravidel jediné proměnné, která byla porušena, a počet hodnot, které porušily každé pravidlo. Také uvádí celkový počet případů porušení pravidel jednotlivých proměnných pro každou proměnnou.
- **Shrnout narušení podle pravidla.** Pro každé pravidlo ověření platnosti jedné proměnné tato volba uvádí proměnné, které porušily pravidlo, a počet neplatných hodnot na proměnnou. Také hlásí celkový počet hodnot, které porušily každé pravidlo napříč proměnnými.

Zobrazte deskriptivní statistiky pro proměnné analýzy. Tato volba vám umožňuje požadovat popisnou statistiku pro proměnné analýzy. Pro každou kategorickou proměnnou je generována tabulka frekvence. Pro proměnné měřítka je vygenerována tabulka souhrnných statistik včetně střední hodnoty, směrodatné odchylky, minima a maxima.

Přesunout případy s narušeními pravidla ověření platnosti na začátek aktivní datové sady. Tato volba přesune případy s jednoduchou proměnnou nebo překročením pravidla proměnné do horní části aktivní datové sady za účelem snadného vypření.

Ověřit uložení dat

Ouško Uložení vám umožňuje uložit proměnné, které zaznamenávají narušení pravidel do aktivní datové sady.

Souhrnné proměnné. To jsou jednotlivé proměnné, které lze uložit. Chcete-li proměnnou uložit, zaškrtněte rámeček. Výchozí názvy pro proměnné jsou k dispozici, můžete je upravit.

- **Prázdný indikátor případu.** Prázdné případy jsou přiřazeny k hodnotě 1. Všechny ostatní případy jsou kódovány 0. Hodnoty proměnné odrážejí rozsah zadaný na kartě Základní kontroly.
- **Kopírovat skupinu ID.** Případy, které mají stejný identifikátor případu (jiné než případy s nekompletními identifikátory), mají přiřazeno stejné číslo skupiny. Případy s jedinečnými nebo neúplnými identifikátory jsou kódovány 0.
- **Neúplný indikátor ID.** Případy s prázdnými nebo neúplnými identifikátory případu mají přiřazenu hodnotu 1. Všechny ostatní případy jsou kódovány 0.
- **Narušení pravidla ověření platnosti.** Jedná se o celkový počet případů porušení pravidel pro ověření jednotlivých proměnných a křížových proměnných pravidel.

Nahradit existující souhrnné proměnné. Proměnné uložené do datového souboru musí mít jedinečné názvy nebo nahradit proměnné se stejným názvem.

Uložit proměnné indikátoru. Tato volba vám umožňuje uložit úplný záznam narušení pravidel ověření platnosti. Každá proměnná odpovídá aplikaci pravidla ověření platnosti a má hodnotu 1, pokud případ poruší pravidlo, a hodnotu 0, pokud tomu tak není.

Automatizovaná příprava dat

Příprava dat pro analýzu je jedním z nejdůležitějších kroků v každém projektu-a tradičně, jeden z nejvíce časově náročné. Automatizovaná příprava dat (automatizovaná příprava dat) zpracovává úkol pro vás, analyzuje data a identifikuje opravy, prozkoumá pole, která jsou problematická, nebo není pravděpodobné, že by mohla být užitečná, podle potřeby odvození nových atributů a zlepšení výkonu pomocí inteligentních screeningových technik. Algoritmus můžete používat plně **automaticky** způsobem, který umožňuje výběr a použití oprav, nebo jej můžete použít v **interaktivním** způsobem, náhled změn před jejich učiněním a přijetím nebo odmítnutím těchto změn, jak chcete.

Pomocí ADP vám umožní rychle a snadno připravit vaše data pro modelové budovy, aniž byste potřebovali předchozí znalosti o souvisejících statistických pojmech. Modely budou mít tendenci vytvářet a skórovat rychleji. Kromě toho díky ADP zlepšuje robustnost automatizovaných procesů modelování.

Poznámka: Když společnost ADP připravuje pole pro analýzu, vytvoří nové pole obsahující úpravy nebo transformace, spíše než nahrazovat existující hodnoty a vlastnosti starého pole. Staré pole se v další analýze nepoužije; jeho role je nastavena na Žádné. Také si všimněte, že žádné informace o hodnotách chybějících uživatelem nejsou přeneseny do těchto nově vytvořených polí a všechny chybějící hodnoty v novém poli chybí systémem.

Příklad. Pojistná společnost s omezenými prostředky na vyšetření pojistných nároků majitele domu chce vytvořit model pro označování podezřelých, potenciálně podvodných nároků. Před sestavením modelu budou připraveny data pro modelování pomocí automatizovaného zpracování dat. Vzhledem k tomu, že chtějí být schopni přezkoumat navrhované transformace dříve, než budou použity transformace, použijí automatizovanou přípravu dat v interaktivním režimu.

Skupina v automobilovém průmyslu uchovává záznamy o prodeji různých osobních motorových vozidel. Ve snaze být schopni identifikovat více než a nedostatečně výkonné modely, chtějí vytvořit vztah mezi prodejem vozidel a charakteristikou vozidla. Budou používat automatizovanou přípravu dat k přípravě dat pro analýzu a sestaví modely s použitím dat "před" a "po", abychom zjistili, jak se výsledky liší.

Jaký je váš cíl? Automatizovaná příprava dat doporučuje kroky přípravy dat, které budou mít vliv na rychlost, s jakou mohou ostatní algoritmy vytvářet modely a zlepšovat prediktivní výkon těchto modelů. To může zahrnovat transformaci, konstrukci a výběr funkcí. Cíl lze také transformovat. Můžete určit priority pro sestavení modelu, na které se má proces přípravy dat soustředit.

- **Vyvážení rychlosti a přesnosti.** Tato volba připravuje data tak, aby měla stejnou prioritu jak rychlosti, se kterými jsou data zpracovávána algoritmem sestavováním modelu, a s přesností předpovědí.
- **Optimalizovat na rychlost.** Tato volba připravuje data tak, aby upřednostňovala rychlost, s jakou jsou data zpracovávána algoritmem sestavováním modelu. Pokud pracujete s velmi velkými datovými sadami, nebo pokud hledáte rychlou odpověď, vyberte tuto volbu.
- **Optimalizujte na přesnost.** Tato volba připravuje data tak, aby upřednostňoval přesnost předpovědí vytvářená algoritmem sestavováním modelů.
- **Vlastní analýza.** Chcete-li ručně změnit algoritmus na kartě Nastavení, vyberte tuto volbu. Všimněte si, že toto nastavení je automaticky vybráno, pokud následně změníte volby na kartě Nastavení, které jsou nekompatibilní s jedním z dalších cílů.

Chcete-li získat automatické přípravy dat

Z nabídky vyberte:

1. Z nabídky vyberte:

   Transformace > **Připravit data pro modelování** > **Automaticky ...**
2. Klepněte na volbu **Spustit**.

Volitelně můžete:

- Zadejte cíl na kartě Cíl.
- Na kartě Pole zadejte přiřazení polí.
- Zadejte odborné nastavení na kartě Nastavení.

Chcete-li získat interaktivní přípravu dat

1. Z nabídky vyberte:

   Transformace > **Připravit data pro modelování** > **Interaktivní ...**
2. Klepněte na tlačítko **Analyzovat** na panelu nástrojů v horní části dialogového okna.
3. Klepněte na kartu Analýza a přezkoumejte navrhované kroky přípravy dat.
4. Jste-li spokojeni, klepněte na tlačítko **Spustit**. Jinak klepněte na volbu **Vymazat analýzu**, změňte veškerá požadovaná nastavení a klepněte na tlačítko **Analyzovat**.

Volitelně můžete:

- Zadejte cíl na kartě Cíl.
- Na kartě Pole zadejte přiřazení polí.
- Zadejte odborné nastavení na kartě Nastavení.
- Chcete-li uložit doporučené kroky přípravy dat do souboru XML, klepněte na volbu **Uložit XML**.

Karta Pole

Karta Pole uvádí, která pole by měla být připravena pro další analýzu.

Použít předdefinované role. Tato volba používá existující informace o poli. Existuje-li jedno pole s rolí jako cíl, bude použito jako cíl; v opačném případě nebude existovat žádný cíl. Jako vstupy budou použity všechny pole s předdefinovanou rolí jako vstup. Je vyžadováno alespoň jedno vstupní pole.

Použít vlastní přiřazení polí. Když přemístíte role polí přesunem polí z jejich výchozích seznamů, dialogové okno se automaticky přepne na tuto volbu. Při vytváření vlastních přiřazení polí zadejte následující pole:

- **Cíl (volitelné).** Plánujete-li sestavení modelů, které vyžadují cíl, vyberte cílové pole. To je podobné nastavení role pole na Cíl.
- **Vstupy.** Vyberte jedno nebo více vstupních polí. To je podobné nastavení role pole na Vstup.

Karta Nastavení

Karta Nastavení se skládá z několika různých skupin nastavení, které můžete upravit k finalizaci způsobu, jakým algoritmus zpracovává vaše data. Provedete-li jakékoli změny výchozích nastavení, která nejsou kompatibilní s ostatními cíli, karta Cíl se automaticky aktualizuje, aby byla vybrána volba **Upravit analýzu** .

Připravit data a časy

Mnoho modelových algoritmů nemůže přímo zpracovat podrobnosti data a času. Tato nastavení vám umožňují odvodit nová data doby trvání, která lze použít jako vstupy modelu z dat a časů ve vašich existujících datech. Pole obsahující data a časy musí být předdefinována s typy úložiště data nebo času. Původní pole data a času se nedoporučují jako přísun modelu po automatizovaném připraveném zpracování dat.

Příprava dat a časů pro modelování. Zrušení výběru této volby zakáže všechny ostatní ovládací prvky Prepare Data & Times při zachování výběru.

Vypočte uplynulý čas do referenčního data. To vytvoří počet rok/měsíců/dní od referenčního data pro každou proměnnou obsahující data.

- **Referenční datum.** Uveďte datum, od kterého bude trvání vypočteno s ohledem na informace o datu ve vstupních datech. Vyberete-li volbu **Dnešní datum** , znamená to, že aktuální systémové datum se vždy použije, když je proveden ADP. Chcete-li použít určité datum, vyberte **Pevné datum** a zadejte požadované datum.
- **Jednotky pro trvání data.** Uveďte, zda by společnost ADP měla automaticky rozhodnout o jednotce doby trvání, nebo vyberte z **Pevné jednotky** roků, měsíců nebo dnů.

Spočítat uplynulou dobu do referenčního času. Tím se vytvoří počet hodin/minutů/sekund od referenčního času pro každou proměnnou obsahující časy.

- **Referenční čas.** Určete čas, od kterého bude trvání vypočítáno s ohledem na informace o čase ve vstupních datech. Vyberete-li **Aktuální čas** , znamená to, že se vždy použije aktuální systémový čas, když se provádí ADP. Chcete-li použít specifický čas, vyberte **Pevný čas** a zadejte požadované podrobnosti.
- **Jednotky pro dobu trvání.** Uveďte, zda by společnost ADP měla automaticky rozhodnout o jednotce doby trvání, nebo vyberte z **Pevné jednotky** hodin, minut nebo sekund.

Extrahovat prvky cyklického času Použijte tato nastavení k rozdělení jednoho pole data nebo času do jednoho nebo více polí. Pokud například vyberete všechna tři políčka data, pole s datem vstupu "1954-05-23" se rozdělí do tří polí: 1954, 5 a 23, přičemž každá z nich bude používat příponu definovanou na panelu **Názvy polí** a původní pole s datem se ignoruje.

- **Extrahovat z dat.** Pro všechny vstupy data určete, zda chcete extrahovat roky, měsíce, dny nebo libovolnou kombinaci.

- **Extrahovat z časů.** U všech časových vstupů určete, zda chcete extrahovat hodiny, minuty, sekundy nebo libovolnou kombinaci.

Vyloučit pole

Špatná kvalita dat může ovlivnit přesnost vašich předpovědí; proto můžete pro vstupní funkce určit přijatelnou úroveň kvality. Všechna pole, která jsou konstantní nebo obsahují 100% chybějící hodnoty, jsou automaticky vyloučena.

Vyloučit vstupní pole nízké kvality. Zrušení výběru této volby vypne při zachování výběru všechny ostatní ovládací prvky Vyloučit pole.

Vyloučit pole s příliš mnoha chybějícími hodnotami. Pole s více než určeným procentem chybějících hodnot jsou odebrány z další analýzy. Uveďte hodnotu větší nebo rovnou 0, která je ekvivalentem zrušení výběru této volby, a menší nebo rovnu 100, ačkoli pole se všemi chybějícími hodnotami jsou automaticky vyloučena. Výchozí hodnota je 50.

Vyloučit nominální pole s příliš mnoha jedinečnými kategoriemi. Jmenovité pole s více než stanoveným počtem kategorií jsou odebrány z další analýzy. Uveďte kladné celé číslo. Výchozí hodnota je 100. To je užitečné pro automatické odebírání polí obsahujících informace o jedinečných záznamech z modelování, jako je ID, adresa nebo název.

Vyloučit kategoriální pole s příliš mnoha hodnotami v jedné kategorii. Pořadová a nominální pole s kategorií, která obsahuje více než uvedené procentní části záznamů, budou odebrány z další analýzy. Uveďte hodnotu větší nebo rovnou 0, která odpovídá zrušení výběru této volby, a menší nebo rovnu 100, ačkoli jsou pole konstant automaticky vyloučena. Předvolba je 95.

Seřídit měření

Upravit úroveň měření. Zrušení výběru této volby zakáže všechny ostatní ovládací prvky Nastavit měření při zachování výběru.

Úroveň měření. Uveďte, zda lze úroveň měření spojitých polí s hodnotami "příliš málo" upravovat na ordinální, a ordinální pole s "příliš mnoha" hodnotami lze upravit tak, aby byla souvislá.

- **Maximální počet hodnot pro ordinální pole.** Ordinální pole s více než určeným počtem kategorií jsou přepracována jako souvislá pole. Uveďte kladné celé číslo. Výchozí hodnota je 10. Tato hodnota musí být větší než nebo rovna minimálnímu počtu hodnot pro souvislá pole.
- **Minimální počet hodnot pro souvislá pole.** Průběžná pole s méně než určeným počtem jedinečných hodnot jsou přepracované jako pořadová pole. Uveďte kladné celé číslo. Výchozí nastavení je 5. Tato hodnota musí být menší než nebo rovna maximálnímu počtu hodnot pro ordinální pole.

Zlepšení kvality dat

Připravte pole pro zlepšení kvality dat. Zrušení výběru této volby zakáže všechny ostatní ovládací prvky Zlepšit kvalitu dat při zachování výběru.

Odlehlá obsluha. Určete, zda mají být nahrazeny odlehlé hodnoty pro vstupy a cíle; pokud ano, zadejte odlehlé kritérium uzavření objektu, měřeno ve směrodatné odchylce a metodu pro nahrazení odlehlých hodnot. Odlehlé hodnoty lze nahradit buď ořezávání (nastavením hodnoty uzavření objektu), nebo jejich nastavením jako chybějící hodnoty. Libovolné odlehlé hodnoty nastavené na chybějící hodnoty odpovídají nastavením zacházení s chybějícím hodnotou vybraným níže.

Nahrazení chybějících hodnot. Uveďte, zda nahradit chybějící hodnoty souvislých, nominálních nebo ordinálních polí.

Přiobjednat nominální pole. Vyberte tuto volbu, chcete-li rekódovat hodnoty nominálního (set) polí od nejmenších (nejméně častých) na největší (nejčastěji se vyskytující) kategorii. Nové hodnoty polí začínají 0 jako nejmenší častou kategorií. Všimněte si, že nové pole bude číselné i v případě, že původní pole je řetězec. Jsou-li například hodnoty dat nominálního pole "A", "A", "A", "B", "C", "C", pak automatizovaná příprava dat by se přeprala "B" do 0, "C" do 1 a "A" do 2.

Změnit měřítko polí

Změna měřítka polí. Zrušení výběru této volby vypne při zachování výběru všechny další ovládací prvky pole Změnit měřítko.

Váha analýzy. Tato proměnná obsahuje váhy analýzy (regrese nebo vzorkování). Váhy analýzy se používají k zohlednění rozdílů v rozptylu v rámci úrovní cílového pole. Vyberte souvislé pole.

Nepřetržitá vstupní pole. Tato akce normalizuje souvislá vstupní pole pomocí transformace **z-skóre** nebo **min/max transformace**. Změna velikosti vstupů je zvláště užitečná, vyberete-li volbu **Provést konstrukci funkcí** v nastavení Vybrat a Vytvořit.

- **Transformace Z-skóre.** Použití pozorovaných středních a směrodatných odchylek jako odhady parametru populace, pole jsou standardizována a pak jsou skóre *z* mapována na odpovídající hodnoty normálního rozdělení s určenými **Finální střední hodnotou** a **Konečnou směrodatnou odchylkou**. Uveďte číslo pro **Finální střední hodnotu** a kladné číslo pro **Konečná směrodatná odchylka**. Výchozí hodnoty jsou 0 a 1, což odpovídá standardizovaným rescaling.
- **Minimální/maximální transformace.** Při použití pozorovaného minima a maxima jako odhadu počtu populačních parametrů jsou pole mapována na odpovídající hodnoty rovnoměrného rozdělení s uvedenými hodnotami **Minimum** a **Maximum**. Uveďte čísla s hodnotou **Maximum** větší než **Minimum**.

Souvislý cíl. Tím se transformuje souvislý cíl pomocí transformace Box-Cox na pole, které má přibližně normální distribuci s určenou hodnotou **Konečná střední hodnota** a **Konečná směrodatná odchylka**. Uveďte číslo pro **Finální střední hodnotu** a kladné číslo pro **Konečná směrodatná odchylka**. Výchozí hodnoty jsou 0 a 1, resp.

Poznámka: Pokud byl cíl transformován ADP, následné modely sestavené pomocí transformovaného cílového skóre transformovaných jednotek. Chcete-li interpretovat a použít výsledky, musíte převést předpokládanou hodnotu zpět na původní měřítko. Další informace naleznete v tématu . Další informace naleznete v tématu “Zpětná hodnocení transformace” na stránce 18 .

Transformační pole

Chcete-li zlepšit prediktivní výkon vašich dat, můžete vstupní pole transformovat.

Transformovat pole pro modelování. Zrušení výběru této volby zakáže všechny ostatní ovládací prvky polí transformace při zachování výběru.

Kategorická vstupní pole K dispozici jsou následující volby:

- **Sloučit řídké kategorie, aby se maximalizovalo přidružení k cíli.** Vyberte tuto volbu, chcete-li vytvořit více parsimonický model tím, že snížíte počet polí, která mají být zpracována ve spojení s cílem. Podobné kategorie jsou identifikovány na základě vztahu mezi vstupem a cílem. Kategorie, které nejsou výrazně odlišné (tj. mající *p*-hodnotu větší než uvedená hodnota), se sloučí. Zadejte hodnotu větší než 0 a menší nebo rovnu 1. Jsou-li všechny kategorie sloučeny do jedné, jsou původní a odvozené verze pole vyloučeny z další analýzy, protože nemají žádnou hodnotu jako prediktor.
- **Když není žádný cíl, slučte řídké kategorie založené na počtech.** Pokud datová sada nemá žádný cíl, můžete zvolit sloučení řídkých kategorií pořadového a nominálního pole. Metoda stejné frekvence se používá ke sloučení kategorií s méně než uvedeným minimálním procentem z celkového počtu záznamů. Uveďte hodnotu větší než nebo rovnou 0 a menší nebo rovnu 100. Výchozí hodnota je 10. Sloučení se zastaví, když nejsou kategorie s méně než uvedeným minimálním procentem případů, nebo když zbývá pouze dvě kategorie.

Nepřetržitá vstupní pole. Pokud datová sada obsahuje kategorický cíl, můžete spolu s silnými přidruženími s silnými přidruženími pokračovat ve zpracování výkonu zpracování. Biny se vytvářejí na základě vlastností "homogenních podskupin", které jsou identifikovány metodou Scheffe použitím specifikované hodnoty *p*-hodnota jako alfa pro kritickou hodnotu pro stanovení homogenní podmnožiny. Zadejte hodnotu větší než 0 a menší nebo rovnu 1. Výchozí hodnota je 0,05. Pokud výsledkem operace binning je pro určité pole jediná přihrádka, jsou původní a binované verze tohoto pole vyloučeny, protože nemají žádnou hodnotu jako prediktor.

Poznámka: Vymírování v ADP se liší od optimálního binnění. Optimální binning používá informace entropie k převedení souvislého pole na kategoriální pole; to potřebuje třídit data a uložit je všechny do paměti. ADP používá homogenní podmnožiny k uložení kontinuálního pole, což znamená, že ADP binning nepotřebuje třídit data a neukládá všechna data v paměti. Použití metody homogenní dílčí sady pro binární pole zásobníku znamená, že počet kategorií po přihrádce je vždy menší než nebo roven počtu kategorií v cíli.

Vyberte a vytvořte

Chcete-li zlepšit prediktivní výkon vašich dat, můžete vytvořit nová pole na základě existujících polí.

Proveďte výběr funkcí. Souvislý vstup je odebrán z analýzy, je-li hodnota *p*-value pro jeho korelaci s cílem vyšší než hodnota zadaná parametrem *p*-value.

Proveďte konstrukci funkcí. Vyberte tuto volbu, chcete-li odvodit nové funkce z kombinace několika existujících funkcí. Staré funkce se v další analýze nepoužívají. Tato volba se používá pouze pro souvislé vstupní funkce, kde je cíl souvislý, nebo kde není žádný cíl.

Názvy polí

Společnost ADP vytváří a používá základní nové názvy, předpony nebo přípony, aby bylo možné snadno identifikovat nové a transformované funkce. Tyto názvy můžete upravit tak, aby byly relevantnější pro vaše vlastní potřeby a data.

Transformovaná a obrazová pole. Zadejte přípony názvů, které se mají použít pro transformovaný cíl a vstupní pole.

Kromě toho zadejte název předpony, který má být použit pro všechny funkce, které jsou konstruovány pomocí nastavení Select a Construct. Nový název se vytvoří připojením číselné přípony k tomuto kořenovému názvu předpony. Formát čísla závisí na tom, kolik nových funkcí je odvozeno, například:

- 1-9 vytvořené funkce budou pojmenovány: feature1 až feature9.
- 10-99 sestavených funkcí bude pojmenováno: feature01 až feature99.
- Bude pojmenováno 100-999 konstruovaných funkcí: feature001 až feature999atd.

Tím je zajištěno, že vybudované funkce budou třídit v rozumném pořadí bez ohledu na to, kolik existuje.

Trvání vypočtených od data a času. Uveďte rozšíření názvu, která se mají použít na trvání vypočtená z data i času.

Cyklické prvky extrahované z dat a časů. Uveďte rozšíření názvu, která se mají použít na cyklické prvky extrahované z dat a časů.

Použití a ukládání transformací

V závislosti na tom, zda používáte dialogová okna Interaktivní nebo Automatická příprava dat, jsou nastavení pro použití a ukládání transformací poněkud odlišná.

Nastavení transformace přípravy interaktivních dat-nastavení transformací

Transformovaná data. Tato nastavení určují, kam se mají uložit transformovaná data.

- **Přidejte nová pole do aktivní datové sady.** Všechna pole vytvořená automatizovanou přípravou dat se přidají jako nová pole do aktivní datové sady. **Aktualizovat role pro analyzované pole** nastaví roli na Žádné pro žádná pole, která jsou vyloučena z další analýzy automatickým přípravou dat.
- **Vytvořte novou datovou sadu nebo soubor obsahující transformovaná data.** Pole doporučená automatizovaným přípravou dat se přidávají do nové datové sady nebo souboru. **Zahrnout neanalyzovaná pole** přidá pole v původní datové sadě, která nebyla uvedena na kartě Pole, do nové datové sady. To je užitečné pro přenos polí obsahujících informace, které se nepoužívají při modelování, jako ID nebo adresa, nebo název, do nové datové sady.

Nastavení automatického použití a uložení dat pro přípravu dat

Skupina Transformovaných dat je stejná jako v Interactive Data Preparation. V modulu Automatická příprava dat jsou k dispozici následující další volby:

Použít transformace. V dialogovém okně Automatická příprava dat zrušením výběru této volby zakážete všechny ostatní ovládací prvky Použít a Uložit při zachování výběru.

Uložit transformace jako syntaxi. Tato volba ukládá doporučené transformace jako syntaxi příkazu do externího souboru. Dialogové okno Příprava interaktivních dat tento ovládací prvek nemá, protože vloží transformace jako syntaxi příkazu do okna syntaxe, pokud klepnete na tlačítko **Vložit**.

Uložit transformace jako XML. Tento příkaz uloží doporučené transformace jako XML do externího souboru, který lze sloučit s modelem PMML pomocí TMS MERGE nebo aplikovaný na jinou datovou sadu pomocí produktu TMS IMPORT. Dialogové okno Příprava interaktivních dat nemá tento ovládací prvek, protože uloží transformace jako XML, pokud klepnete na tlačítko **Uložit XML** v panelu nástrojů v horní části dialogového okna.

Karta Analýza

Poznámka: Karta Analýza se používá v dialogovém okně Příprava interaktivních dat a umožňuje vám zkontrolovat doporučené transformace. Dialogové okno Automatické přípravy dat neobsahuje tento krok.

1. Když jste spokojeni s nastavením ADP, včetně změn provedených na kartách Cíl, Pole a Nastavení, klepněte na **Analyzovat data**; algoritmus použije nastavení na datové vstupy a zobrazí výsledky na kartě Analýza.

Karta Analýza obsahuje jak tabulkový, tak grafický výstup, který shrnuje zpracování vašich dat a zobrazuje doporučení, jak mohou být data modifikována nebo zlepšena pro přidělení skóre. Poté můžete přezkoumat a buď přijmout, nebo odmítnout tato doporučení.

Karta Analýza se skládá ze dvou panelů, z hlavního pohledu vlevo a z propojeného a pomocného zobrazení vpravo. K dispozici jsou tři hlavní pohledy:

- Souhrn zpracování polí (výchozí). Další informace naleznete v tématu “Souhrn zpracování polí ” na stránce 12 .
- Pole. Další informace naleznete v tématu “Pole ” na stránce 13 .
- Souhrn akcí. Další informace naleznete v tématu “Souhrn akcí ” na stránce 14 .

Existují čtyři propojené/pomocné pohledy:

- Prediktivní napájení (výchozí). Další informace naleznete v tématu “Predictive Power ” na stránce 14 .
- Tabulka polí. Další informace naleznete v tématu “Tabulka polí ” na stránce 14 .
- Podrobnosti pole. Další informace naleznete v tématu “Podrobnosti pole ” na stránce 15 .
- Podrobnosti akce. Další informace naleznete v tématu “Podrobnosti akce ” na stránce 16 .

Odkazy mezi zobrazeními

V hlavním pohledu je podtržený text v tabulkách řízen zobrazením v propojeném pohledu. Klepnutí na text vám umožní získat podrobnosti o konkrétním poli, sadě polí nebo kroku zpracování. Spoj, který jste naposledy vybral, se zobrazí tmavší barvou; pomáhá identifikovat spojení mezi obsahem obou panelů zobrazení.

Obnova zobrazení

Chcete-li znovu zobrazit původní doporučení analýzy a zrušit všechny provedené změny v pohledech Analýza, klepněte na tlačítko **Resetovat** v dolní části hlavního panelu pohledu.

Souhrn zpracování polí

Tabulka Souhrn zpracování polí poskytuje snímek předpokládaného celkového dopadu zpracování, včetně změn stavu funkcí a počtu sestavených funkcí.

Všimněte si, že žádný model není ve skutečnosti sestaven, takže neexistuje žádné měřítko nebo graf změn celkové prediktivní schopnosti před a po přípravě dat; místo toho můžete zobrazit grafy prediktivní moci jednotlivých doporučených prediktorů.

V tabulce jsou zobrazeny následující informace:

- Počet cílových polí.
- Počet původních (vstupních) prediktorů.
- Prediktory se doporučují pro použití v analýze a modelování. To zahrnuje celkový počet doporučených polí; počet původních, netransformovaných, doporučených polí; počet doporučených transformovaných polí (s výjimkou intermediačních verzí polí, polí odvozených z prediktorů data a času a konstruovaných prediktorů); počet doporučených polí, která jsou odvozena z polí data a času; a počet doporučených prediktorů.
- Počet vstupních prediktorů není doporučen pro použití v žádné formě, ať už v původní podobě, jako odvozené pole, nebo jako vstup sestaveného prediktoru.

Je-li libovolné informace **Pole** podtržené, můžete klepnutím zobrazit více podrobností v propojeném pohledu. Podrobnosti o polích **Cíl**, **Vstupní funkce**a **Vstupní funkce nejsou použity** jsou zobrazeny v zobrazení propojeného tabulky polí. Další informace naleznete v tématu “Tabulka polí ” na stránce 14 . **Funkce doporučené pro použití v analýze** jsou zobrazeny v zobrazení propojeného prediktivního napájení. Další informace naleznete v tématu “Predictive Power ” na stránce 14 .

Pole

Hlavní zobrazení polí zobrazuje zpracovaná pole a to, zda společnost ADP doporučuje je používat v následných modelech. Můžete potlačit doporučení pro jakékoli pole; například k vyloučení sestavených funkcí nebo zahrnutých funkcí, které společnost ADP doporučuje vyloučit. Pokud bylo pole transformováno, můžete rozhodnout, zda přijmout navrhovanou transformaci, nebo použít původní verzi.

Pohled Pole se skládá ze dvou tabulek, jednoho pro cíl a jednoho pro prediktory, které byly buď zpracovány, nebo vytvořeny.

Cílová tabulka

Tabulka **Cíl** se zobrazí pouze v případě, že je cíl definován v datech.

Tabulka obsahuje dva sloupce:

- **Název.** Jedná se o název nebo popisek cílového pole; původní název je vždy použit, i když bylo pole transformováno.
- **Úroveň měření.** Zobrazí se ikona reprezentující úroveň měření; přesuňte ukazatel myši nad ikonu pro zobrazení popisku (souvislý, ordinální, nominální atd.), který popisuje data.

  Pokud byl cíl transformován, sloupec **Úroveň měření** odráží konečnou transformovanou verzi. *Poznámka*: Nemůžete vypnout transformace pro cíl.

Tabulka predikátů

Tabulka **Prediktory** se vždy zobrazí. Každý řádek tabulky představuje pole. Při výchozím nastavení jsou řádky řazeny v sestupném pořadí prediktivního výkonu.

Pro běžné funkce je původní název vždy použit jako název řádku. Původní i odvozené verze polí s datem/časem se objevují v tabulce (v samostatných řádcích); tabulka také obsahuje sestavené prediktory.

Všimněte si, že transformované verze polí zobrazené v tabulce vždy představují konečné verze.

Ve výchozím nastavení jsou v tabulce predikátů zobrazena pouze doporučená pole. Chcete-li zobrazit zbývající pole, vyberte pole **Zahrnout nedoporučená pole v tabulce** nad tabulkou; tato pole se poté zobrazí v dolní části tabulky.

Tabulka obsahuje následující sloupce:

- **Verze, která se má použít.** Zobrazí se rozevírací seznam, který řídí, zda bude pole použito po směru toku a zda má být použita navrhovaná transformace. Ve výchozím nastavení tento rozevírací seznam odráží doporučení.

  Pro běžné prediktory, které byly transformovány, má rozevírací seznam tři volby: **Transformované**, **Původní**a **Nepoužívat**.

  Pro netransformované běžné prediktory jsou volby následující: **Původní** a **Nepoužívat**.

  Pro odvozená pole data/času a konstruované prediktory jsou tyto volby: **Transformované** a **Nepoužít**.

  Pro původní datová pole je rozevírací seznam zakázán a nastaven na hodnotu **Nepoužít**.

  Poznámka: Pro prediktory s původní i transformovanou verzí automaticky mění verze **Původní** a **Transformované** automaticky nastavení **Úroveň měření** a **Prediktivní napájení** pro tyto funkce.
- **Název.** Název každého pole je odkaz. Po klepnutí na název se zobrazí další informace o poli v propojeném pohledu. Další informace naleznete v tématu “Podrobnosti pole ” na stránce 15 .
- **Úroveň měření.** Zobrazí se ikona představující datový typ; podržením ukazatele myši nad ikonou se zobrazí popisek (souvislý, ordinální, nominální atd.), který popisuje data.
- **Prediktivní napájení.** Prediktivní napájení se zobrazuje pouze pro pole, která společnost ADP doporučuje. Tento sloupec se nezobrazí, pokud není definován žádný cíl. Prediktivní rozsah výkonu je v rozsahu 0 až 1, přičemž větší hodnoty označují, že "lepší" prediktory. Obecně platí, že prediktivní výkon je užitečný při porovnávání prediktorů v rámci analýzy ADP, ale prediktivní hodnoty výkonu by neměly být porovnávány mezi analýzami.

Souhrn akcí

Pro každou akci provedené automatizovanou přípravou dat se vstupní prediktory transformují a/nebo odfiltrují; pole, která přežijí jednu akci, se použijí v další. Pole, která přežijí do posledního kroku, se pak doporučuje používat při modelování, zatímco vstupy k transformovaným a konstruovaným predictorům jsou odfiltrovány.

Souhrn akcí je jednoduchou tabulkou, která uvádí akce zpracování prováděné ADP. Kde je **Akce** podtržená, klepnutím zobrazte další podrobnosti v propojeném zobrazení o provedených akcích. Další informace naleznete v tématu “Podrobnosti akce ” na stránce 16 .

Poznámka: Zobrazeny jsou pouze původní a konečné transformované verze každého pole, nikoli mezilehlé verze, které byly použity během analýzy.

Predictive Power

Zobrazí se standardně při prvním spuštění analýzy nebo při výběru **Prediktorů doporučených pro použití v analýze** v hlavním pohledu Souhrn zpracování polí, v grafu se zobrazí prediktivní výkon doporučených prediktorů. Pole jsou seřazena podle prediktivního výkonu, s polem s nejvyšší hodnotou, která se objevuje nahoře.

V případě transformovaných verzí běžných predikátů název pole odráží vaši volbu přípony v panelu Názvy polí na kartě Nastavení; například: *_transformed*.

Ikony úrovně měření se zobrazují za názvy jednotlivých polí.

Prediktivní výkon každého doporučeného prediktoru je vypočítán buď z lineární regrese, nebo z modelu Bayes, v závislosti na tom, zda je cíl spojitý nebo kategorický.

Tabulka polí

Zobrazí se, když klepnete na volbu **Cíl**, **Prediktory**nebo **Prediktory nepoužité** v hlavním zobrazení Souhrn zpracování polí, v zobrazení tabulky polí se zobrazí jednoduchá tabulka se seznamem příslušných funkcí.

Tabulka obsahuje dva sloupce:

- **Název.** Název prediktoru.

  Pro cíle je použit původní název nebo popisek pole, i když byl cíl transformován.

V případě transformovaných verzí běžných predikátů název odráží vaši volbu přípony v panelu Názvy polí na kartě Nastavení; například: *_transformed*.

U polí odvozených z dat a časů se použije název konečné transformované verze, například: *bdate_years*.

Pro konstruované prediktory se používá název konstruovaného prediktoru; například: *Predictor1*.

- **Úroveň měření.** Zobrazí se ikona reprezentující datový typ.

  Pro cíl **Úroveň měření** vždy odráží transformovanou verzi (pokud byl cíl transformován); například změna z ordinálního (seřazené sady) na souvislý (rozsah, měřítko) nebo naopak.

Podrobnosti pole

Zobrazí se, když klepnete na libovolné **Název** v hlavním pohledu Pole, v pohledu Podrobnosti o poli jsou distribuce, chybějící hodnoty a prediktivní grafy výkonu (jsou-li použitelné) pro vybrané pole. Kromě toho se zobrazí také historie zpracování pole a název transformovaného pole (je-li k dispozici).

Pro každou sadu grafů jsou dvě verze zobrazeny vedle sebe, aby bylo možné porovnat pole s použitou a bez použití transformací; pokud transformovaná verze pole neexistuje, graf se zobrazí pouze pro původní verzi. Pro odvozené pole data nebo času a konstruované prediktory se grafy zobrazují pouze pro nový prediktor.

Poznámka: Je-li pole vyloučeno z důvodu použití příliš mnoha kategorií, zobrazí se pouze historie zpracování.

Distribuční graf

Rozdělování souvislých polí je zobrazeno jako histogram, s křivkou normální křivky a svislým vztažným řádkem pro střední hodnotu; kategorická pole jsou zobrazena jako pruhový graf.

Histogramy jsou označeny tak, že ukazují směrodatnou odchylku a šikmost, avšak šikmost se nezobrazí, pokud je počet hodnot 2 nebo méně nebo odchylka původního pole je menší než 10-20.

Přesunutím ukazatele myši nad graf se zobrazí buď průměr pro histogramy, nebo počet a procentní část celkového počtu záznamů pro kategorie v pruhových grafech.

Graf chybějících hodnot

Výsečové grafy porovnávají procentní podíl chybějících hodnot s použitými transformacemi a bez použití transformací; popisky grafu zobrazují procentní část.

Pokud společnost ADP provedla chybějící zpracování hodnot, pak výsečový graf po transformaci také obsahuje hodnotu náhrady jako popisek -- tj. použitou hodnotu místo chybějících hodnot.

Podržte ukazatel myši nad grafem, abyste zobrazili počet chybějících hodnot a procento celkového počtu záznamů.

Prediktivní napájecí graf

Pro doporučená pole zobrazují pruhové grafy prediktivního napájení před a po transformaci. Pokud byl cíl transformován, vypočítá se vypočtená předpokládaná mocnost pro transformovaný cíl.

Poznámka: Prediktivní grafy výkonu se nezobrazí, pokud není definován žádný cíl, nebo pokud na něj klepnete v hlavním panelu pohledu.

Přesunutím ukazatele myši nad graf zobrazíte hodnotu prediktivního výkonu.

Zpracování tabulky historie

Tabulka ukazuje, jak byla odvozena transformovaná verze pole. Akce prováděné ADP jsou uvedeny v pořadí, ve kterém byly provedeny; avšak u některých kroků bylo možné provést více akcí pro určité pole.

Poznámka: Tato tabulka se nezobrazí pro pole, která nebyla transformována.

Informace v tabulce jsou rozděleny do dvou nebo tří sloupců:

- **Akce.** Název akce. Například Spojité prediktory. Další informace naleznete v tématu “Podrobnosti akce ” na stránce 16 .
- **Podrobnosti.** Seznam provedených zpracování. Například Transformovat na standardní jednotky.
- **Funkce.** Zobrazuje se pouze pro sestavené prediktory, zobrazuje lineární kombinaci vstupních polí, například .06*age + 1.21*height.

Podrobnosti akce

Zobrazí se, když vyberete podtržené **Akce** v hlavním zobrazení Souhrn akcí, pohled Podrobnosti akce zobrazí jak akce specifické pro danou akci, tak běžné informace pro každý krok zpracování, který byl proveden; podrobnosti specifické pro danou akci se zobrazí jako první.

Pro každou akci se popis použije jako nadpis v horní části propojeného pohledu. Podrobnosti specifické pro akci jsou zobrazeny pod titulkem a mohou obsahovat podrobnosti o počtu odvozených prediktorů, přepracované pole, cílové transformace, kategorie sloučené nebo přeskupené a prediktory konstruované nebo vyloučené.

Při zpracování jednotlivých akcí se může počet prediktorů použitých ve zpracování změnit, například protože prediktory jsou vyloučeny nebo sloučeny.

Poznámka: Pokud byla akce vypnuta nebo nebyl zadán žádný cíl, zobrazí se při klepnutí na tlačítko Souhrn akcí v hlavním pohledu Souhrn akcí chybová zpráva jako místo podrobností akce.

Existuje devět možných akcí; avšak ne všechny jsou nutně aktivní pro každou analýzu.

Tabulka textových polí

V tabulce se zobrazí počet položek:

- Prediktory byly vyloučeny z analýzy.

Tabulka predikátů data a času

V tabulce se zobrazí počet položek:

- Doby trvání odvozené od prediktorů data a času.
- Prvky data a času.
- Odvozené prediktory data a času, celkem.

Referenční datum nebo čas se zobrazí jako poznámka pod čarou, pokud bylo vypočteno jakékoli datum trvání.

Tabulka screeningu prediktoru

V tabulce se zobrazí počet následujících predikátů, které byly vyloučeny ze zpracování:

- Konstanty.
- Prediktory s příliš mnoha chybějícími hodnotami.
- Prediktory s příliš mnoha případy v jedné kategorii.
- Nominální pole (sady) s příliš mnoha kategoriemi.
- Předpověřovatelé podrobeni detekční kontrole, celkem.

Zkontrolovat tabulku úrovně měření

V tabulce se uvádí počet přepracovaných polí, rozdělených do následujících čísel:

- Ordinální pole (řazené sady) přepracované jako souvislá pole.
- Průběžná pole přepracované jako pořadová pole.
- Celkový počet přepracování.

Pokud nebyla žádná vstupní pole (cíl nebo prediktory) spojité nebo ordinální, zobrazí se jako poznámka pod čarou.

Tabulka odlehlých

Tabulka zobrazuje počty, jak byly obslouženy všechny outliery.

- Počet souvislých polí, pro které byly nalezeny a oříznuty odlehlé hodnoty, nebo počet souvislých polí, pro které byly nalezeny a nastaveny přelehlé pole, v závislosti na vašich nastaveních na panelu Příprava vstupů a cíle na kartě Nastavení.
- Počet souvislých polí vyloučených z toho důvodu, že byly konstantní, po odlehlé hodnotě.

Jedna poznámka pod čarou ukazuje odlehlější hodnotu uzavření; zatímco jiná poznámka pod čarou se zobrazí, pokud nebyla průběžná žádná vstupní pole (cíl nebo prediktory).

Tabulka chybějících hodnot

V tabulce se zobrazí počet polí, u kterých byly chybějící hodnoty nahrazeny, rozdělené do:

- Cíl. Tento řádek se nezobrazuje, není-li uveden žádný cíl.
- Prediktory. To je dále rozděleno do počtu jmenovitých (set), ordinal (seřazených sad) a souvislých.
- Celkový počet chybějících chybějících hodnot.

Cílová tabulka

V tabulce se zobrazí, zda byl cíl transformován, zobrazen jako:

- Transformace Box-Cox na normality. Tento stav se dále dělí na sloupce, které zobrazují uvedená kritéria (střední a směrodatná odchylka) a Lambda.
- Cílové kategorie znovu přiobjednají ke zlepšení stability.

Tabulka Kategorických predikátů

V tabulce se zobrazí počet kategorických prediktorů:

- jejichž kategorie byly přeskupeny od nejnižší po nejvyšší, aby se zlepšila stabilita.
- jejichž kategorie byly sloučeny za účelem maximalizace přidružení k cíli.
- jejichž kategorie byly sloučeny pro zpracování řídkých kategorií.
- Vyloučeno z důvodu nízkého přidružení k cíli.
- Vyloučeny, protože byly po sloučení konstantní.

Pokud neexistovaly žádné kategoriální prediktory, zobrazí se poznámka pod čarou.

Tabulka souvislých predikátů

Existují dvě tabulky. První zobrazuje jeden z následujících počtu transformací:

- Hodnoty predikátu transformované na standardní jednotky. Kromě toho se zobrazuje počet transformovaných prediktorů, určený střední hodnota a směrodatná odchylka.
- Hodnoty predikátu mapované na společný rozsah. Kromě toho se zobrazuje počet predikátů transformovaných pomocí transformace min-max, stejně jako uvedené minimální a maximální hodnoty.
- Hodnoty prediktoru binned a počet prediktorů binned.

Druhá tabulka zobrazuje podrobnosti o konstrukci prostoru prediktoru, které se zobrazují jako počet prediktorů:

- Vyrobeno.
- Vyloučeno z důvodu nízkého přidružení k cíli.
- Vyloučeny, protože byly konstantní po binning.
- Vyloučeny, protože byly po konstrukci konstantní.

Zobrazí se poznámka pod čarou, pokud nebyly nalezeny žádné souvislé prediktory.

Zpětná hodnocení transformace

Pokud byl cíl transformován ADP, následné modely sestavené pomocí transformovaného cílového skóre transformovaných jednotek. Chcete-li interpretovat a použít výsledky, musíte převést předpokládanou hodnotu zpět na původní měřítko.

1. Chcete-li zpětně transformovat skóre, z nabídek vyberte:

   Transformace > **Připravit data pro modelování** > **Skóre zpětné transformace ...**
2. Vyberte pole pro zpětný převod. Toto pole by mělo obsahovat modelované hodnoty transformovaného cíle, které jsou předpovězeny modelem.
3. Uveďte příponu pro nové pole. Toto nové pole bude obsahovat předpovídané hodnoty modelu v původním měřítku netransformovaného cíle.
4. Uveďte umístění souboru XML, který obsahuje transformace ADP. Mělo by se jednat o soubor uložený z dialogových oken pro interaktivní nebo automatickou přípravu dat. Další informace naleznete v tématu “Použití a ukládání transformací ” na stránce 11 .

Identifikace neobvyklých případů

Postup detekce anomálií je určen pro neobvyklé případy založené na odchylkách od norem jejich skupin. Procedura je navržena tak, aby rychle zjišťovala neobvyklé případy pro účely sledování dat v kroku analýzy dat, a to před jakoukoli inferenciální analýzou dat. Tento algoritmus je určen pro detekci generických anomálií; to znamená, že definice nenormálního případu není specifická pro žádnou konkrétní aplikaci, jako je například detekce neobvyklých platebních vzorců v odvětví zdravotní péče nebo zjišťování praní špinavých peněz ve finančním odvětví, v němž lze definici anomálie dobře definovat.

Příklad. Analytik dat najímán k sestavování prediktivních modelů pro výsledky léčebných úhozů se zabývá kvalitou dat, protože takové modely mohou být citlivé na neobvyklé pozorování. Některé z těchto odlehlých pozorování představují skutečně jedinečné případy a jsou tudíž nevhodné pro predikci, zatímco jiné pozorování jsou způsobeny chybami vstupu dat, ve kterých jsou hodnoty technicky "správné", a proto nemohou být zachyceny postupy validace dat. Procedura Identifikace neobvyklých případů vyhledá a ohlásí tyto odlehlé hodnoty, aby mohl analytik rozhodnout o tom, jak je zpracovat.

Statistika. Tento postup vytváří skupiny rovnocenných uzlů, skupinové standardy rovnocenných uzlů pro spojité a kategoriální proměnné, indexy anomálií založené na odchylkách od norem skupiny rovnocenných uzlů a proměnné dopadu proměnných pro proměnné, které nejvíce přispívají k případu, který je považován za neobvyklý.

Aspekty dat

Data. Tato procedura pracuje se spojitými a kategoriálními proměnnými. Každý řádek představuje zřetelné pozorování a každý sloupec představuje odlišnou proměnnou, na níž jsou založeny skupiny rovnocenných uzlů. V datovém souboru může být k dispozici identifikační proměnná případu pro označení výstupu, ale v analýze se nepoužije. Chybějící hodnoty jsou povoleny. Proměnná váhy, je-li zadána, je ignorována.

Model detekce lze použít na nový testovací datový soubor. Prvky zkušebních údajů musí být stejné jako prvky údajů o výcviku. A v závislosti na nastavení algoritmu může být zpracování chybějící hodnoty použité k vytvoření modelu použito pro testovací datový soubor před bodováním.

Pořadí případů. Všimněte si, že řešení může záviset na pořadí případů. Chcete-li minimalizovat efekty objednávky, náhodně objednejte případy. Chcete-li ověřit stabilitu daného řešení, možná budete chtít získat několik různých řešení s případy setříděnými v různých náhodných objednávkách. V situacích s extrémně velkými velikostmi souborů lze provést více spuštění se vzorkem případů seřazených v různých náhodných příkazech.

Předpoklady. Algoritmus předpokládá, že všechny proměnné jsou nekonstantní a nezávislé a že žádný případ nemá chybějící hodnoty pro žádnou vstupní proměnnou. Každá spojitá proměnná se předpokládá, že má normální (Gaussovu) distribuci, a každá kategorická proměnná se předpokládá, že má polynomiální distribuci. Empirické vnitřní testování ukazuje, že tento postup je poměrně spolehlivý na porušování

předpokladů nezávislosti a distribuční předpoklady, ale je si vědom toho, jak dobře jsou tyto předpoklady splněny.

identifikace neobvyklých případů

1. Z nabídky vyberte:

   Data > **Identifikovat neobvyklé případy ...**
2. Vyberte alespoň jednu proměnnou analýzy.
3. Volitelně můžete vybrat proměnnou identifikátoru případu, která se má použít při popisování výstupu.

Pole s neznámou úrovní měření

Výstraha na úrovni měření se zobrazí, když je úroveň měření pro jednu nebo více proměnných (polí) v datové sadě neznámá. Jelikož úroveň měření ovlivňuje výpočet výsledků pro tuto proceduru, všechny proměnné musí mít definovanou úroveň měření.

Data skenování. Přečte data v aktivní datové sadě a přiřadí výchozí úroveň měření k jakýmkoli polím s momentálně neznámou úrovní měření. Je-li datová sada velká, může to nějakou dobu trvat.

Přiřadit ručně. Otevře dialogové okno se seznamem všech polí s neznámou úrovní měření. Toto dialogové okno můžete použít k přiřazení úrovně měření k těmto polím. Úroveň měření můžete také přiřadit v pohledu Proměnné v editoru dat.

Vzhledem k tomu, že úroveň měření je pro tuto proceduru důležitá, nemůžete přistupovat k dialogovému oknu pro spuštění této procedury, dokud nebude mít všechna pole definovanou úroveň měření.

Identifikace neobvyklých výstupních případů

Seznam neobvyklých případů a důvodů, proč jsou považovány za neobvyklé. Tato volba vytvoří tři tabulky:

- Seznam indexů případu anomálie zobrazuje případy, které jsou identifikovány jako neobvyklé a zobrazují jejich odpovídající hodnoty indexu anomálie.
- Seznam ID rovnocenných uzlů anomálií zobrazuje neobvyklé případy a informace týkající se jejich odpovídajících skupin rovnocenných uzlů.
- Seznam příčin anomálií zobrazuje číslo případu, proměnnou příčiny, hodnotu proměnné dopadu, hodnotu proměnné a normu proměnné z jednotlivých příčin.

Všechny tabulky jsou řazeny podle indexu anomálií v sestupném pořadí. Kromě toho se ID případů zobrazí, pokud je proměnná identifikátoru případu uvedena na kartě Proměnné.

Souhrny. Ovládací prvky v této skupině vytvářejí souhrny distribuce.

- **Normy stejné skupiny.** Tato volba zobrazí tabulku s proměnlivou normou proměnných (pokud se v analýze používá libovolná souvislá proměnná) a tabulka norem kategorií kategoriálních proměnných (pokud je v analýze použita libovolná kategorická proměnná). V tabulce se spojitými proměnnými se zobrazí střední a směrodatná odchylka každé souvislé proměnné pro každou rovnocennou skupinu. Tabulka standardů kategorických proměnných zobrazuje režim (nejoblíbenější kategorie), frekvenci a procentní část četnosti každé kategoriální proměnné pro každou skupinu rovnocenných uzlů. Jako hodnoty standardních hodnot v analýze se používají střední hodnoty proměnné a režim kategorické proměnné.
- **Anomálie indices.** Souhrn indexu anomálií zobrazuje deskriptivní statistiky pro index anomálie v případech, které jsou identifikovány jako nejobvyklejší.
- **Příčina výskytu podle proměnné analýzy.** Z každého důvodu tabulka zobrazuje frekvenci a frekvenci četnosti výskytu jednotlivých proměnných jako důvod. V tabulce jsou také uvedeny deskriptivní statistiky vlivu jednotlivých proměnných. Je-li maximální počet důvodů nastaven na 0 na kartě Volby, tato volba není k dispozici.
- **Zpracované případy.** Souhrn zpracování případu zobrazuje počty a procentní části počtu pro všechny případy v aktivní datové sadě, případy zahrnuté a vyloučené v analýze a případy v každé skupině rovnocenných uzlů.

Identifikovat případy, kdy nejsou běžné případy

Uložit proměnné. Ovládací prvky v této skupině umožňují uložit proměnné modelu do aktivní datové sady. Můžete se také rozhodnout nahradit existující proměnné, jejichž názvy jsou v konfliktu s proměnnými, které mají být uloženy.

- **Anomálie index.** Uloží hodnotu indexu anomálie pro každý případ na proměnnou s určeným názvem.
- **Skupiny rovnocenných uzlů.** Uloží ID skupiny rovnocenných uzlů, počet případů a velikost jako procentní část pro každý případ k proměnným s uvedeným názvem rootname. Je-li například zadán parametr rootname *Peer* , vygenerují se proměnné *Peerid*, *PeerSize*a *PeerPctSize* . *Peerid* je ID skupiny rovnocenných uzlů případu, *PeerSize* je velikost skupiny a *PeerPctSize* je velikost skupiny jako procento.
- **Důvody.** Uloží sady proměnných zdůvodnění s uvedeným rootname. Sada proměnných odůvodnění se skládá z názvu proměnné jako příčiny, jeho proměnného dopadu dopadu, jeho vlastní hodnoty a hodnoty standardu. Počet sad závisí na počtu požadovaných důvodů na kartě Volby. Je-li například zadán parametr rootname *Reason* , vygenerují se proměnné *ReasonVar_k*, *ReasonMeasure_k*, *ReasonValue_k*a *ReasonNorm_k* , kde *k* je *k*. důvodem. Tato volba není k dispozici, pokud je počet důvodů nastaven na 0.

Exportovat soubor modelu. Umožňuje vám uložit model ve formátu XML.

Identifikace neobvyklých případů Chybějící hodnoty

Karta Chybějící hodnoty se používá k ovládání obsluhy chybějících a systémových hodnot chybějících systémem.

- **Vyloučit chybějící hodnoty z analýzy.** Případy s chybějícími hodnotami jsou z analýzy vyloučeny.
- **Zahrnout chybějící hodnoty do analýzy.** Chybějící hodnoty spojitých proměnných jsou nahrazeny odpovídajícími velkými prostředky a chybějící kategorie kategorických proměnných jsou seskupeny a jsou považovány za platnou kategorii. Zpracované proměnné se pak použijí v analýze. Volitelně můžete požádat o vytvoření další proměnné, která představuje proporci chybějících proměnných v každém případě a tuto proměnnou v analýze použijte.

Identifikace neobvyklých voleb případů

Kritéria pro identifikaci neobvyklých případů. Tyto výběry určují, kolik případů je zahrnuto do seznamu anomálií.

- **Procentní část případů s nejvyšší hodnotou indexu anomálií.** Uveďte kladné číslo, které je menší než nebo rovno 100.
- **Pevný počet případů s nejvyšší hodnotou indexu anomálií.** Uveďte kladné celé číslo, které je menší než nebo rovno celkovému počtu případů v aktivní datové sadě, které jsou použité v analýze.
- **Identifikujte pouze případy, jejichž hodnota indexu anomálie odpovídá nebo překračuje minimální hodnotu.** Uveďte nezáporné číslo. Případ je považován za anomální, je-li jeho hodnota indexu anomálie větší než nebo rovna zadanému bodu uzavření. Tato volba se používá společně s volbami **Procentní část případů** a **Pevný počet případů** . Pokud například uvedete pevný počet 50 případů a hodnota uzavření 2, seznam anomálií se bude skládat nanejvýš z 50 případů, každý s hodnotou indexu anomálie, která je větší než nebo rovna 2.

Počet rovnocenných skupin. Procedura bude hledat nejlepší počet rovnocenných skupin mezi zadanými minimálními a maximálními hodnotami. Hodnoty musí být kladná celá čísla a minimální hodnota nesmí překročit maximum. Jsou-li zadané hodnoty shodné, procedura předpokládá pevný počet rovnocenných skupin.

Poznámka: V závislosti na množství variací ve vašich datech mohou nastat situace, kdy počet rovnocenných skupin, které data může podporovat, je menší než počet uvedený jako minimum. V takové situaci může procedura vytvořit menší počet rovnocenných skupin.

Maximální počet důvodů. Důvod se skládá z variabilního ukazatele dopadu, názvu proměnné z tohoto důvodu, hodnoty proměnné a hodnoty odpovídající skupiny rovnocenných uzlů. Uveďte nezáporné celé

číslo; pokud se tato hodnota rovná nebo překračuje počet zpracovaných proměnných, které se použijí v analýze, jsou zobrazeny všechny proměnné.

Další funkce příkazu DETECTANOMALY

Jazyk syntaxe příkazu vám také umožňuje:

- Vynechte několik proměnných v aktivní datové sadě z analýzy bez explicitního určení všech proměnných analýzy (pomocí dílčího příkazu EXCEPT).
- Určete nastavení pro vyvážení vlivu průběžných a kategoriálních proměnných (pomocí klíčového slova MLWEIGHT na dílčím příkazu CRITERIA).

Úplné informace o syntaxi najdete v příručce *Command Syntax Reference* .

Optimální ukotvení

Procedura Optimal Breinning diskretizuje jednu nebo více proměnných (označovaných napříště jako **binning input variables**) rozrozdělením hodnot každé proměnné do přihrádek. Tvorba koše je optimální s ohledem na proměnnou kategoriálního průvodce, která "dohlíží" na binning procesu. Místo původních hodnot dat lze pro další analýzu použít místo původních hodnot dat.

Příklady. Snížení počtu odlišných hodnot, které má proměnná, má řadu použití, včetně:

- Požadavky na údaje z jiných postupů. Diskretizované proměnné lze považovat za kategorické pro použití v procedurách, které vyžadují kategoriální proměnné. Například procedura kontingenční tabulky vyžaduje, aby všechny proměnné byly kategorické.
- Ochrana dat. Nahlašování sloučených hodnot místo skutečných hodnot může pomoci zajistit ochranu soukromí vašich zdrojů dat. Optimální vypalovací postup může vést k výběru zásobníků.
- Rychlost. Některé postupy jsou efektivnější při práci se sníženým počtem odlišených hodnot. Například při použití diskretizovaných proměnných lze zlepšit rychlost Multinomial Logistic Regression.
- Odhalování úplného nebo kvaziúplného oddělení údajů.

Optimální versus vizuální ukotvení. Dialogová okna Vizuální vychycení nabízí několik automatických metod pro vytváření přihrádek bez použití směrné proměnné. Tato pravidla "bez dozoru" jsou užitečná pro tvorbu popisných statistik, jako jsou frekvenční tabulky, ale Optimal Binning je nadřazený, když váš konečný cíl je produkovat prediktivní model.

Výstup. Postup vytváří tabulky bodů pro kolekce pro kolekce a deskriptivní statistiky pro každou vstupní proměnnou binning. Kromě toho můžete uložit nové proměnné do aktivní datové sady obsahující binované hodnoty vstupních proměnných binning a uložit pravidla binning jako syntaxi příkazu pro použití v diskretizaci nových dat.

Aspekty optimálního ukotvení dat

Data. Tento postup očekává, že vstupní proměnné binning budou nastaveny na měřítko, číselné proměnné. Tato vodicí proměnná by měla být kategorická a může být řetězcová nebo číselná.

Chcete-li dosáhnout optimálního sváření

1. Z nabídky vyberte:

   Transformace > **Optimální ukotvení ...**
2. Vyberte jednu nebo více vstupních proměnných binning.
3. Vyberte vodicí proměnnou.

Proměnné obsahující hodnoty binned data nejsou standardně generovány. Použijte kartu Uložit k uložení těchto proměnných.

Optimálně vypalovací výstup

Karta Výstup řídí zobrazení výsledků.

- **Koncové body pro přihrádky.** Zobrazí sadu koncových bodů pro každou vstupní proměnnou binning.
- **Deskriptivní statistiky pro proměnné, které jsou binované.** Pro každou vstupní proměnnou binning tato volba zobrazí počet případů s platnými hodnotami, počet případů s chybějícími hodnotami, počet odlišných platných hodnot a minimální a maximální hodnoty. Pro proměnnou vodítka tato volba zobrazí distribuci třídy pro každou související vstupní proměnnou binning.
- **Model entropie pro proměnné, které jsou binované.** Pro každou vstupní proměnnou binning tato volba zobrazí ukazatel prediktivní přesnosti proměnné s ohledem na směrnou proměnnou.

Optimální ukládání do neaktivního stavu

Uložit proměnné do aktivní datové sady. Proměnné obsahující hodnoty binned data lze použít místo původních proměnných v další analýze.

Uložit ukotvení pravidel jako syntaxe. Generuje syntaxi příkazu, která může být použita pro binární soubor dat. Pravidla kódování jsou založena na bodech omezení určených algoritmem binning.

Optimální vychycení chybějících hodnot

Karta Chybějící hodnoty uvádí, zda se s chybějícími hodnotami zachází pomocí lisu nebo odstranění po dvojicích. Uživatel-chybějící hodnoty jsou vždy považovány za neplatné. Při převodu hodnot původní proměnné do nové proměnné se uživatelem chybějící hodnoty převedou na systém-chybí.

- **Peirwise.** Tato volba pracuje na každé dvojici vodicí proměnné a binning vstupní proměnné. Procedura použije všechny případy s nechybějícími hodnotami na vstupní proměnné průvodce a binning.
- **Listwise** Tato volba funguje napříč všemi proměnnými zadanými na kartě Proměnné. Pokud u případu chybí nějaká proměnná, je celý případ vyloučen.

Optimální volby ukotvení

Předběžné zpracování. "Předinebinning" binning input variables with many distinct values can improve processing time without a great utratit in the quality of the final bins. Maximální počet přihrádek dává horní hranici počtu vytvořených přihrádek. Pokud tedy uvedete 1000 jako maximum, ale vstupní proměnná binning má méně než 1000 odlišných hodnot, počet předem zpracovaných přihrádek vytvořených pro vstupní proměnnou binning se bude rovnat počtu odlišených hodnot ve vstupní proměnné binning.

Volně naplněné Bins. Procedura může příležitostně produkovat zásobníky s velmi malými případy. Následující strategie odstraní tyto pseudoškrtání:

Pro danou proměnnou předpokládejme, že algoritmus nalezl n_{final} výřezů, a tedy $n_{final}+1$. Pro zásobníky $i = 2, ..., n_{final}$ (druhý nejnižší podhodnota v koši s druhou nejvyšší hodnotou), výpočet

$$\frac{sizeof(b_i)}{\min(sizeof(b_{i-1}), sizeof(b_{i+1}))}$$

kde *sizeof (b)* je počet případů v přihrádce.

Je-li tato hodnota nižší než uvedená prahová hodnota sloučení, produkt b_n se považuje za řídce naplněný a je sloučen s b_{i-1} nebo b_{i+1}, podle toho, co má nižší informace o třídě entropii.

Procedura provede jediné předání v přihrádkách.

Koncové body zásobníku. Tato volba uvádí, jak je definován dolní limit intervalu. Vzhledem k tomu, že tento postup automaticky určuje hodnoty výřezů, je to z velké části věc preference.

První (nejnižší)/Poslední (nejvyšší) zásobník. Tyto volby určují, jak jsou definovány minimální a maximální počet bodů omezení pro každou vstupní proměnnou binning. Obecně platí, že procedura předpokládá, že vstupní proměnné binning mohou mít libovolnou hodnotu na řádku reálného čísla, ale pokud máte nějaký teoretický nebo praktický důvod omezení rozsahu, můžete jej svázat s nejnižšími/nejvyššími hodnotami.

Další funkce příkazu OPTIMAL BINNING

Jazyk syntaxe příkazu vám také umožňuje:

- Proveďte nedozorované spojení přes metodu rovnoměrných kmitočtů (pomocí dílčího příkazu `CRITERIA`).

Úplné informace o syntaxi najdete v příručce *Command Syntax Reference* .

Upozornění

Tyto informace byly vytvořeny pro produkty a služby poskytované v USA. Tento materiál může být dostupný od IBM v jiných jazycích. K povolení přístupu však může být vyžadováno vlastnictví kopie produktu nebo verze produktu v tomto jazyce.

Společnost IBM nemusí nabízet produkty, služby nebo funkce uvedené v tomto dokumentu v jiných zemích. Informace o produktech a službách, které jsou aktuálně k dispozici ve vaší oblasti, získáte od lokálního zástupce společnosti IBM. Odkazy na produkty, programy nebo služby společnosti IBM neuvádí ani neimplikují, ze lze použít pouze daný produkt, program nebo službu společnosti IBM. Lze použít libovolný funkčně ekvivalentní produkt, program nebo službu neporušující práva duševního vlastnictví společnosti IBM. Vyhodnocení a ověření funkčnosti produktů, programů nebo služeb, které nepatří společnosti IBM, je však zodpovědností uživatele.

Společnost IBM může vlastnit patenty nebo nevyřízené žádosti o patenty zahrnující předměty popsané v tomto dokumentu. Vlastnictví tohoto dokumentu neposkytuje licenci k těmto patentům. Dotazy na licence můžete písemně odeslat na následující adresu:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
USA

S dotazy na licence týkající se dvoubajtových informací (DBCS) se obraťte na oddělení intelektuálního vlastnictví společnosti IBM v dané zemi, nebo je odešlete písemně na následující adresu:

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

SPOLEČNOST INTERNATIONAL BUSINESS MACHINES CORPORATION POSKYTUJE TUTO PUBLIKACI "TAKOVOU, JAKÁ JE", BEZ JAKÝCHKOLIV ZÁRUK, VYJÁDŘENÝCH NEBO ODVOZENÝCH VČETNĚ, MIMO JINÉ, ODVOZENÝCH ZÁRUK NEPORUŠENÍ PRÁV TŘETÍCH STRAN, ZÁRUKY PRODEJNOSTI NEBO VHODNOSTI PRO URČITÝ ÚČEL. Některé právní řády u určitých transakcí nepřipouštějí vyloučení záruk výslovně vyjádřených nebo vyplývajících z okolností, a proto se na vás výše uvedené omezení nemusí vztahovat, a proto se vás toto prohlášení nemusí týkat.

Uvedené údaje mohou obsahovat technické nepřesnosti nebo typografické chyby. Údaje zde uvedené jsou pravidelně upravovány a tyto změny budou zahrnuty v nových vydáních této publikace. Společnost IBM může kdykoli bez upozornění provádět vylepšení nebo změny v produktech či programech popsaných v této publikaci.

Jakékoliv odkazy v této publikaci na webové stránky jiných společností nežIBM jsou poskytovány pouze pro pohodlí uživatele a nemohou být žádným způsobem vykládány jako doporučení těchto webových stránek. Materiály uvedené na těchto webových stránkách nejsou součástí materiálů pro tento produkt IBM a použití uvedených stránek je pouze na vlastní nebezpečí.

IBM může použít nebo distribuovat jakékoli informace, které jí poskytnete, libovolným způsobem, který společnost považuje za odpovídající, bez vzniku jakýchkoliv závazků vůči vám.

Vlastníci licence k tomuto programu, kteří chtějí získat informace o možnostech (i) výměny informací s nezávisle vytvořenými programy a jinými programy (včetně tohoto) a (ii) oboustranného využití vyměňovaných informací, mohou kontaktovat informační středisko na adrese:

IBM Director of Licensing
IBM Corporation

North Castle Drive, MD-NC119
Armonk, NY 10504-1785
USA

Poskytnutí takových informací může být podmíněno dodržením určitých podmínek a požadavků zahrnujících v některých případech uhrazení stanoveného poplatku.

Licencovaný program popsaný v tomto dokumentu a veškerý licencovaný materiál k němu dostupný jsou společností IBM poskytovány na základě podmínek uvedených ve smlouvách IBM Customer Agreement, IBM International Program License Agreement nebo v jiné ekvivalentní smlouvě.

Citovaná data o výkonu a příklady klienta jsou uvedeny pouze pro názornost. Skutečné výsledky výkonu se mohou lišit v závislosti na specifických konfiguracích a provozních podmínkách.

Informace týkající se produktů jiných společností nežIBM byly získány od dodavatelů těchto produktů, z jejich publikovaných sdělení, nebo z jiných veřejně dostupných zdrojů. IBM tyto produkty netestovala a nemůže potvrdit přesnost údajů o výkonu, kompatibilitě nebo jiná tvrzení týkající se produktů jiných společností nežIBM. Otázky týkající se možností produktů jiných společností nežIBM by měly být adresovány dodavatelům těchto produktů.

Prohlášení týkající se budoucího směru vývoje nebo záměrů společnosti IBMse mohou změnit nebo mohou být zrušena bez předchozího upozornění a představují pouze cíle a záměry.

Tyto údaje obsahují příklady dat a sestav používaných v běžných obchodních operacích. Aby byla představa úplná, používají se v příkladech jména osob, společností, značek a produktů. Všechna tato jména jsou fiktivní a jakákoliv podobnost se skutečnými lidmi nebo obchodními podniky je čistě náhodná.

COPYRIGHT - LICENCE:

Tyto informace obsahují ukázkové aplikační programy ve zdrojovém jazyku a ilustrují různé programovací techniky na různých operačních platformách. Tyto ukázkové programy můžete bez závazků vůči společnosti IBM jakýmkoli způsobem kopírovat, měnit a distribuovat za účelem vývoje, používání, odbytu či distribuce aplikačních programů odpovídajících rozhraní API pro operační platformu, pro kterou byly ukázkové programy napsány. Tyto příklady nebyly důkladně testovány ve všech podmínkách. Společnost IBM proto nemůže zaručit spolehlivost, upotřebitelnost nebo funkčnost těchto programů. Ukázkové programy jsou poskytovány "JAK JSOU", bez záruky jakéhokoli druhu. IBM nenese odpovědnost za žádné škody vzniklé ve spojení s Vaším užíváním ukázkových programů.

Jakákoli kopie nebo část těchto ukázkových programů nebo jakékoli odvozené dílo musí obsahovat následující poznámku o autorských právech:

Ochranné známky

IBM, logo IBM a ibm.com jsou ochranné známky nebo registrované ochranné známky společnosti International Business Machines Corp., registrované v mnoha jurisdikcích po celém světě. Ostatní názvy produktů a služeb mohou být ochrannými známkami společnosti IBM nebo jiných společností. Aktuální seznam ochranných známek společnosti IBM je k dispozici na webu na stránce "Copyright and trademark information" na adrese www.ibm.com/legal/copytrade.shtml.

Adobe, logo Adobe, PostScript a logo PostScript jsou buď registrované ochranné známky, nebo ochranné známky společnosti Adobe Systems Incorporated ve Spojených státech anebo v dalších zemích.

Intel, logo Intel, Intel Inside, logo Intel Inside, Intel Centrino, logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium a Pentium jsou ochranné známky nebo registrované ochranné známky společnosti Intel Corporation nebo jejích dceřiných společností ve Spojených státech a případně v dalších jiných zemích.

Linux je registrovaná ochranná známka Linuse Torvaldse ve Spojených státech a případně v dalších jiných zemích.

Microsoft, Windows, Windows NT a logo Windows jsou ochranné známky společnosti Microsoft Corporation ve Spojených státech a případně v dalších jiných zemích.

UNIX je registrovaná ochranná známka společnosti The Open Group ve Spojených státech a případně v dalších jiných zemích.

Java a všechny ochranné známky a loga založené na jazyce Java jsou ochranné známky nebo registrované ochranné známky společnosti Oracle anebo příbuzných společností.

Rejstřík

A

Automatická příprava dat 6
automatizovaná příprava dat
    analýza pole 13
    backtransformování skóre 18
    Cíle 6
    konstrukce funkcí 11
    normalizovat souvislý cíl 10
    obnovení zobrazení 12
    odkazy mezi zobrazeními 12
    Podrobnosti akce 16
    podrobnosti o poli 15
    pole 8
    pole názvu 11
    pole ve stupních měřítka 10
    použít transformace 11
    prediktivní výkon 14
    příprava data a času 8
    Souhrn akcí 14
    souhrn zpracování polí 12
    tabulka polí 14
    transformační pole 10
    upravit úroveň měření 9
    Výběr funkcí 11
    vyloučení polí 9
    zlepšení kvality dat 9
    Zobrazení modelu 12

B

binning bez dozoru
    versus pečení pod dohledem 21

C

cyklické časové prvky
    automatizovaná příprava dat 8

D

Definovat pravidla ověření platnosti
    pravidla pro křížovou proměnnou 3
    pravidla s jednou proměnnou 2
dozorování pod dohledem
    v optimálním vyzrňování 21
    versus nedozorovaný binning 21
duplicitní identifikátory případu
    v Ověřit data 6
důvody
    v identifikaci neobvyklých případů 19, 20

CH

chybějící hodnoty
    v identifikaci neobvyklých případů 20

I

Identifikace neobvyklých případů
    chybějící hodnoty 20
    soubor modelu exportu 20
    ukládání proměnných 20
    volby 20
    výstup 19
indexy anomálie
    v identifikaci neobvyklých případů 19, 20

K

koncové body pro přihrádky
    v optimálním vyzrňování 21
konstrukce funkcí
    v automatizovaném zpracování dat 11

M

MDLP
    v optimálním vyzrňování 21

N

narušení pravidla ověření platnosti
    v Ověřit data 6
neúplné identifikátory případu
    v Ověřit data 6
normalizovat souvislý cíl 10

O

Optimální ukotvení
    chybějící hodnoty 22
    uložit 22
    volby 22
    výstup 21
Ověřit data
    pravidla pro křížovou proměnnou 5
    pravidla s jednou proměnnou 5
    ukládání proměnných 6
    výstup 5
    základní kontroly 4
ověřování dat
    v Ověřit data 3

P

porušení pravidel pro ověření platnosti
    v Ověřit data 6
pravidla ověření napříč proměnnými
    v definici pravidel pro ověření platnosti 3
    v Ověřit data 5
pravidla ověření platnosti 1
pravidla ověření platnosti s jednou proměnnou

pravidla ověření platnosti s jednou proměnnou *(pokračování)*
v definici pravidel pro ověření platnosti 2
v Ověřit data 5
pravidla přihrádkování
v optimálním vyzrňování 22
prázdné případy
v Ověřit data 6
pre-binning
v optimálním vyzrňování 22
Příprava interaktivních dat 6

S

skupiny rovnocenných uzlů
v identifikaci neobvyklých případů 19, 20

T

Transformace Box-Cox
v automatizovaném zpracování dat 10

V

váha analýzy
v automatizovaném zpracování dat 10
Výběr funkcí
v automatizovaném zpracování dat 11
výpočet doby trvání
automatizovaná příprava dat 8
výpočetní trvání
automatizovaná příprava dat 8

Z

Zobrazení modelu
v automatizovaném zpracování dat 12