Zrozumienie danych – przykładowe techniki umożliwiające zrozumienie danych[1]

Podczas prowadzenia projektu eksploracji danych po fazie inicjalnej – zrozumienia celów i założeń biznesowych następuje faza zrozumienia danych. Faza ta łączy ze sobą: identyfikację źródeł danych, opisanie danych oraz analizę pozyskanych danych, również pod kątem ich jakości. W znacznej mierze polega ona na przeprowadzeniu eksploracyjnej analizy danych (EDA).

W kolejnych punktach zostaną przedstawione techniki wykonania różnych zadań dotyczących etapu zrozumienia danych z wykorzystaniem narzędzia KNIME ver 4.1.1.

I. Analiza ilościowa danych

Jednym z zadań jest analiza ilościowa danych. Analiza ilościowa danych zostanie pokazana na przykładzie zbioru danych o markach samochodów.

Zbiór danych zapisanych w pliku *cars.txt* zawiera informację o 261 markach samochodów wyprodukowanych w latach siedemdziesiątych i osiemdziesiątych, atrybuty mówią o zużyciu paliwa (*mpg - miles per gallon*), liczbie cylindrów (*cylinders*), pojemności silnika w calach sześciennych (*cubicinches*), liczbie koni mechanicznych (*hp - horse power*), wadze auta (*weightlbs* - waga w funtach), czasie przyspieszania do 60 mil na godzinę *(timeto-60* - w sekundach), roku debiutu na rynku (*year*) oraz kraju pochodzenia (*brand*).

1. Po uruchomieniu systemu i wybraniu odpowiedniej przestrzeni roboczej wybierz z repozytorium węzłów (ang. *Node Repository*) węzeł *IO→Read→File Reader* i przenieś go na przestrzeń roboczą projektu. Opis wybranego węzła pokaże się w zakładce *Node Description*.
2. Pod prawym przyciskiem myszy po kliknięciu na wybrany węzeł znajduje się menu kontekstowe. Wybierz *Configure...*
3. W polu *Enter ASCII Data File Location* wybierz plik *cars.txt*. Obejrzyj pozostałe opcje i zastosuj zdefiniowane ustawienia.
4. Dołóż do projektu węzeł *Views→Local (Swing)→Interactive Table*. Połącz jego wejście z wyjściem węzła *File Reader*. Zapoznaj się z opisem węzła.
5. Z menu kontekstowego węzła wykonaj polecenie *Execute and Open Views*. Obejrzyj przykłady w tabeli.

[1] Przykłady zostały opracowane na podstawie: Kursu IBM: Introduction to IBM SPSS Modele rand Data Mining (Student Guide) oraz książki Daniela T. Larose „Odkrywanie wiedzy z danych" Wprowadzenie do eksploracji danych. Wydawnictwo Naukowe PWN, Warszawa 2006.

6. Przenieś na przestrzeń roboczą projektu węzeł *Analytics→Statistics→Statistics* i połącz jego wejście do wyjścia węzła *File Reader*. Zapoznaj się z opisem węzła *Statistics*.
7. Obejrzyj wartości pokazywane przez węzeł *Statistics* (opcja *Execute and Open Views*).

II. Analiza brakujących wartości

Brakujące dane są zawsze problemem przy tworzeniu rozwiązań eksploracji danych. Należy zidentyfikować brakujące dane, a następnie podjąć decyzję co z tymi danymi, a właściwie ich brakiem zrobić.

Brakujące dane można podzielić na kilka klas: brak wartości, pusty łańcuch znaków albo wartość ustalona przez administratora. Pierwsze dwa typy brakujących wartości są zazwyczaj wychwytywane na etapie odczytywania danych. Ostatni z typów brakujących danych można wykryć używając metod identyfikacji punktów oddalonych (to zagadnienie jest opisane w punkcie III, przy identyfikacji punktów oddalonych).

Identyfikacja brakujących danych zostanie pokazana na przykładzie zbioru danych z informacjami demograficznymi.

Zbiór danych zapisanych w pliku *SmallSampleMissing.txt* zawiera informacje demograficzne o poszczególnych osobach identyfikowanych pewnym identyfikatorem – ID. Atrybuty mówią o wieku - AGE, płci - SEX, typie regionu, w jakim zamieszkuje dana osoba - REGION, przychodzie - INCOME, stanie cywilnym MARRIED, liczbie dzieci CHILDREN i posiadaniu samochodu - CAR.

1. Wybierz z repozytorium węzłów (ang. Node Repository) węzeł *IO→Read→File Reader* i przenieś go na przestrzeń roboczą projektu.
2. Pod prawym przyciskiem myszy po kliknięciu na wybrany węzeł znajduje się menu kontekstowe. Wybierz *Configure...*
3. W polu *Enter ASCII Data File Location* wybierz plik *SmallSampleMissing.txt*. Obejrzyj pozostałe opcje i zastosuj zdefiniowane ustawienia.
4. Do obsługi brakujących danych służy narzędzie *Manipulation→Column→Transform→Missing Value*. Przenieś węzeł *Missing Value* na przestrzeń roboczą projektu i połącz wyjście węzła *File Reader* z wejściem węzła *Missing Value*. Zapoznaj się z opisem węzła *Missing Value*.
5. Poeksperymentuj z różnymi ustawieniami węzła *Missing Value*.
6. Skonfiguruj węzeł *Missing Value* (posłuż się zakładką *Column Settings* - poszczególne atrybuty) tak, aby uzupełniał brakujące wartości atrybutu *Income* jako średnią ze znanych wartości, a wartości atrybutu *Sex* na wartość dominującą. Zapisz plik z

uzupełnionymi danymi i sprawdź, statystyki (węzeł *Statistics*) dla obu plików. Czy wartości brakujące zostały przez ten węzeł wykryte?

III. Identyfikacja punktów (obserwacji) oddalonych

Punkty (obserwacje) oddalone są skrajnymi wartościami, które znajdują się blisko granic zakresu danych lub są sprzeczne z ogólnym trendem pozostały danych.

Punkty oddalone mogą:
- być błędnymi danymi,
- wartościami brakującymi,
- powodować błędy w pewnych metodach statystycznych wrażliwych na punkty oddalone.

Najbardziej popularne narzędzia wykorzystywane do identyfikacji punktów oddalonych to:
- histogramy
- wykresy punktowe rozproszone (ang. *scatter plot*)
- wykresy pudełkowe (ang. *box plot*)

Do identyfikacji punktów oddalonych zostanie wykorzystany zbiór z informacją o klientach formy telekomunikacyjnej.

W pliku *churn.txt* znajduje się zbiór danych składający się z 20 zmiennych informujących o 3333 klientach, razem ze wskazaniem, czy zrezygnowali z usług firmy (zmienna *churn*). Zmienne są następujące: - stan (*state*) – 50 stanów i Dystrykt Kolumbia, - czas współpracy (*account length*) – czas posiadania konta, - kod *(area code*) – kod obszaru, - telefon (*phone*) – telefon, - plan mędzynarodowy (*intl plan*) – czy klient przystąpił do planu międzynarodowego, - poczta głosowa (*vmail plan*) – czy klient przystąpił do planu poczty głosowej, - liczba wiadomości (*vmail message*) – liczba wiadomości w poczcie głosowej, - dzień minut (*day mins*) – liczba minut, które klient zużył w ciągu dnia, - dzień rozmowy (*day calls*) – liczba połączeń w dzień, - dzień opłata (*day charge*) – całkowita opłata za rozmowy w dzień, - wieczór minuty (*eve mins*) – całkowita liczba minut wieczorem, - wieczór rozmowy (*eve calls*) – liczba połączeń wieczorem, - wieczór opłata (*eve charge*) – całkowita opłata za rozmowy wieczorem, - noc minuty (*night mins*) – całkowita liczba minut w nocy, - noc rozmowy (*night calls*) – liczba połączeń w nocy, - noc opłata (*night charge*) – całkowita opłata za rozmowy w nocy, - międzynarodowe minuty (*intl mins*) – całkowita liczba minut na połączenia międzynarodowe, - międzynarodowe rozmowy (*intl calls*) – liczba połączeń międzynarodowych, - międzynarodowe opłaty (*intl charge*) – całkowita opłata za rozmowy w połączeniach międzynarodowych - liczba rozmów z BOK (*custServ calls*) – liczba połączeń z biurem obsługi klienta.

1. Wybierz z repozytorium węzłów (ang. *Node Repository*) węzeł *IO→Read→File Reader* i przenieś go na przestrzeń roboczą projektu.
2. Pod prawym przyciskiem myszy po kliknięciu na wybrany węzeł znajduje się menu kontekstowe. Wybierz *Configure...*
3. W polu *Enter ASCII Data File Location* wybierz plik *churn.txt*. Obejrzyj pozostałe opcje i zastosuj zdefiniowane ustawienia.
4. Przenieś węzeł *Views→Local→Box Plot* na przestrzeń roboczą projektu i połącz wyjście węzła *File Reader* z wejściem węzła *Box Plot*. Zapoznaj się z opisem węzła *Box Plot*.
5. Wybierając z menu kontekstowego węzła *Box Plot* opcję *Execute and Open Views* obejrzyj dla poszczególnych kolumn kwartyle i rozstęp międzykwartylowy.
6. Usuń wszystkie kolumny poza *VMail Message*; zwróć uwagę na punkt oddalony dla kolumny *VMail Message*.

IV. Zadania samodzielne

Zadania samodzielne należy wykonać dla pliku *churn.txt*.

1. Przeanalizuj zależności pomiędzy poszczególnymi zmiennymi.
2. Zidentyfikuj punkty oddalone wynikające z wartości zmiennej Day Calls.