Niektórzy najlepiej uczą się na przykładach.
Tak samo jak niektóre maszyny.
Marcin Szeliga

Rozdział 2

Praca z SQL Server Machine Learning Services

W tym rozdziale podaję dwa przykłady zastosowania SQL Server Machine Learning Services do zaawansowanej analizy danych. Czytelnik nie znajdzie w nim ani słowa na temat teorii uczenia maszynowego. Tak jak w poprzednim, w tym **rozdziale omawiam narzędzia, a nie techniki uczenia maszynowego**. Różnica między rozdziałami polega na sposobie przedstawiania tych samych zagadnień – tym razem SQL Server Machine Learning Services zostało użyte do rozwiązania dwóch typowych problemów.

2.1. Wykrywanie oszustw

Wykrywanie oszustw było jednym z pierwszych praktycznych zastosowań zaawansowanej analityki. My też swoją przygodę z SQL Server Machine Learning Services zaczniemy od wykrycia prób oszustwa. Naszym zadaniem będzie wytypowanie firm wystawiających naszemu zleceniodawcy fałszywe faktury. Jedyne czym dysponujemy, to dane o fakturach wystawionych zleceniodawcy.

```
SELECT *
FROM [BenfordFraud].[Invoices];
```

```
VendorNumber     VoucherNumber     CheckNumber       InvoiceNumber     InvoiceDate
PaymentDate      DueDate  InvoiceAmount     PONumber
387707  62286    56229    8016     2005-07-19        2005-08-18        2005-08-18
```

```
9844.46 9872
387707  61574    56157    9384     2005-10-27        2005-12-02        2005-11-26
4614.70 6598
387707  97399    59740    8732     2005-09-12        2005-10-15        2005-10-12
2569.51 4045
387707  97680    59768    5769     2005-02-07        2005-02-28        2005-03-09
4185.07 6230
…
```

Każda faktura jest opisana w osobnym wierszu tabeli `[BenfordFraud].[Invoices]`; tabela liczy 245 830 wierszy. Chociaż danych mamy sporo, wykrycie w nich podejrzanych firm wymaga kreatywności i przeprowadzenia zaawansowanej analizy.

Spróbujmy porównać faktyczny rozkład wartości wybranej zmiennej z jej spodziewanym rozkładem. W określeniu spodziewanego rozkładu pomoże nam znajomość rozkładu Benforda. Opisuje on rozkład prawdopodobieństwa występowania pierwszych cyfr w wielu różnych danych statystycznych. Zgodne z rozkładem Benforda są m.in. częstości występowania pierwszych cyfr w rocznikach statystycznych i stałych fizycznych, długościach rzek, wielkościach miast, liczbie mieszkańców krajów itp.

Fenomen ten po raz pierwszy (w 1881 r.) opisał astronom Simon Newcomb [11]. Zauważył on, że pierwsze strony tablic logarytmicznych są bardziej wyświechtane, co sugerowałoby, że są częściej używane. Fakt występowania tego rozkładu w obserwowanych danych został potwierdzony w 1938 r. przez fizyka Franka Benforda. Po zbadaniu danych z 20 różnych obszarów sformułował on prawo, nazwane od jego nazwiska prawem Benforda [12]: **jeżeli zmienna może przyjmować różne rzędy wielkości, prawdopodobieństwo wystąpienia jej pierwszej cyfry *k* wynosi**

$$\log_{10}\left(1+\frac{1}{k}\right)$$

Rozkład Benforda został pokazany na rysunku 2.1.

Obecnie prawo Benforda jest powszechnie używane do wykrywania oszustw podatkowych. Spróbujmy go użyć do sprawdzenia wiarygodności kwot, na jakie poszczególne firmy wystawiły faktury.

Będziemy potrzebować danych o częstości występowania pierwszych cyfr w kwotach wystawionych faktur, czyli takich danych, jakie zawraca funkcja `[BenfordFraud].[VendorInvoiceDigits]`:

```
SELECT *
FROM [BenfordFraud].[VendorInvoiceDigits] (default)
ORDER BY VendorNumber;
VendorNumber     Digits   Freq
105313  1.00     173
```