Eksploracja danych

Ćwiczenia nr 10

Plik dane_telco.csv

W pliku *dane_telco.csv* znajdują się dane dotyczące wysiłków firmy telekomunikacyjnej na rzecz zmniejszenia odpływu jej abonentów. Każdy rekord odpowiada jednemu klientowi i zawiera m.in. następujące jego atrybuty:

- *longmon* – daleki zasięg w poprzednim miesiącu (koszt),
- *tollmon* – bezpłatne połączenia w zeszłym miesiącu (koszt),
- *equipmon* – wynajem sprzętu w zeszłym miesiącu (koszt),
- *cardmon* – karta telefoniczna w zeszłym miesiącu (koszt),
- *wiremon* – usługi bezprzewodowe w zeszłym miesiącu (koszt),
- *multiline* – wiele linii: 0 – nie, 1 – tak,
- *voice* – poczta głosowa: 0 – nie, 1 – tak,
- *pager* – usługa przywoływania: 0 – nie, 1 – tak,
- *internet*: 0 – nie, 1 – tak,
- *callid* – identyfikacja numeru: 0 – nie, 1 – tak,
- *callwait* – połączenie oczekujące: 0 – nie, 1 – tak,
- *forward* – przekazywanie połączeń: 0 – nie, 1 – tak,
- *confer* – telekonferencja: 0 – nie, 1 – tak,
- *ebill* – elektroniczna faktura: 0 – nie, 1 – tak.

Operator ten planuje zsegmentować bazę klientów na podstawie schematów korzystania z jego usług. Jeżeli uda się to zrobić, firma będzie mogła proponować klientom bardziej atrakcyjne pakiety.

a. Zbadaj zależność zmiennych zawierających informację o kosztach usług z ostatniego miesiąca (końcówka nazwy *mon*) i kosztach usług od początku umowy (końcówka nazwy *ten*). Czy jest sens budować model w oparciu o oba typy zmiennych?

b. Zbadaj kształt rozkładu wymienionych wyżej zmiennych (np. przy użyciu wykresów skrzynkowych, histogramów) a następnie przygotuj je do dalszej analizy przekształcając je następująco:

    i. zmienne jakościowe ustandaryzuj,
    ii. zmienne ilościowe przekształć funkcją logarytmiczną (lub ln(x+1)), a następnie ustandaryzuj.

c. Narysuj kilka różnych trójwymiarowych wykresów rozrzutu dla wybranych zmiennych otrzymanych w poprzednim punkcie. Czy udaje Ci się zaobserwować naturalne skupiska danych?

d. Na podstawie przekształconych zmiennych pogrupuj klientów tak, aby firmie łatwo było dotrzeć z atrakcyjnymi ofertami do poszczególnych sektorów abonentów. Uruchom algorytm k-średnich, ustalając najpierw *k=3*.

e. Zastanów się, czy otrzymany podział na grupy jest zadowalający i posłuży do osiągnięcia celów firmy. Jeżeli nie, znajdź rozwiązanie, które to umożliwi.

f. Spróbuj opisać otrzymane grupy i nadać im adekwatne nazwy.
g. Sprawdź jaki jest rozkład zmiennej *churn* w każdej z otrzymanych grup.