Eksploracja danych

Ćwiczenia nr 5

Plik *churn_pl_dane.sav* zawiera dane o abonentach pewnej amerykańskiej firmy telekomunikacyjnej. Kolejne zmienne oznaczają:

- *Telefon* – numer telefonu abonenta (zmienna służąca do identyfikacji klienta),
- *Czas współpracy* – czas współpracy w miesiącach,
- *Liczba wiadomości* – liczba wiadomości w poczcie głosowej,
- *Dzień (Wieczór, Noc, Międzynarodowe) minuty* – przeciętna miesięczna liczba minut rozmów prowadzonych w taryfie Dzień (Wieczór, Noc, Międzynarodowe),
- *Dzień (Wieczór, Noc, Międzynarodowe) rozmowy* – miesięczna liczba rozmów w taryfie Dzień (Wieczór, Noc, Międzynarodowe),
- *Dzień (Wieczór, Noc, Międzynarodowe) opłata* – przeciętna miesięczna opłata w taryfie Dzień (Wieczór, Noc, Międzynarodowe),
- *Liczba rozmów z BOK* – liczba rozmów z biurem obsługi klienta,
- *Planmiedzy01* – czy klient ma aktywowaną usługę planu międzynarodowego (0 – nie, 1 – tak),
- *Pocztagl01* – czy klient ma aktywowaną usługę poczty głosowej (0 – nie, 1 – tak),
- *Rezygnacja* – czy klient zrezygnował z usług firmy (0 – nie, 1 – tak).

Wykonaj następujące polecenia:

a) Wczytaj plik *churn_pl_dane.sav.*

b) Sprawdź, czy w danych występują zmienne silnie skorelowane.

c) Ustaw generator liczb losowych na wartość 123. Następnie zbuduj metodą drzew CRT model, który będzie klasyfikował klientów ze względu na możliwość rezygnacji z usług firmy. Jako predyktorów użyj zmiennych *Czaswspółpracy, Liczbawiadomości, Dzieńminuty, Dzieńrozmowy, Wieczórminuty, Wieczórrozmowy, Nocminuty, Nocrozmowy, Międzynarodoweminuty, Międzynarodowerozmowy, LiczbarozmówzBOK*, *Planmiedzy01*. Pamiętaj o podziale danych na zbiór uczący i testowy. Ustaw maksymalną głębokość drzewa na 5, minimalną liczbę elementów w węźle na 10, a minimalną liczbę elementów w liściu na 5. Zapisz przewidywaną wartość i przewidywane prawdopodobieństwo. Wyznacz ważność predyktorów dla modelu i narysuj jej wykres. Które predyktory model uznał za najważniejsze?

d) Obejrzyj strukturę drzewa i postaraj się ją zinterpretować. Co decyduje o rezygnacji klienta z usług firmy? Oceń jakość otrzymanego modelu. Porównaj uzyskane na zbiorze uczącym i testowym wartości trafności, czułości i specyficzności. Narysuj krzywe ROC dla danych ze zbiory testowego i policz pole pod nią.

e) Zbuduj model drzewa raz jeszcze, przycinając je w celu uniknięcia przeuczenia. Sprawdź, co stało się z wartościami trafności, czułości i specyficzności. Czy otrzymałeś teraz lepszy model?

f) Zbuduj jeszcze raz model CRT, ale zastosuj teraz równe prawdopodobieństwa a priori dla wszystkich grup. Porównaj otrzymane modele (początkowy, przycięty, z równymi prawdopodobieństwami a priori). Który ma najwyższą czułość, który trafność, a który specyficzność?

g) Zastosuj najlepszy z uzyskanych modeli na pliku *churn_pl_dane_nowe.sav*. Porównaj wartości otrzymane przez model z prawdziwymi (zmienna *Churn*).

h) Usuń z listy predyktorów zmienne odnoszące się do rozmów prowadzonych w nocy, a następnie spróbuj zbudować i zinterpretować model szacujący liczbę minut rozmów prowadzonych przez klienta w taryfie nocnej.