Laboratorium 5

Bartlomiej Karaban

20 03 2018

Wymagane biblioteki

```
library(data.table)
library(rpart)
library(rpart.plot)
library(randomForest)
library(caret)
```

Zadanie 1

Wykonaj samodzielnie (bez używania dedykowanych pakietów/funkcji) wykres ROC. Wejściem będzie wektor prawdopodobieństw i wektor oznaczeń good/bad.

Zadanie 2

- Zdefiniuj zmienną celu `IfPayment6M` jako pojawienie się wpłaty w pierwszych 6ciu miesięcach obsługi dla każdego CaseId. Poza nowo utworzoną zmienną zachowaj tylko i wyłącznie dane z tabeli cases.
- Następnie podziel zbiór danych przy użyciu funkcji `createDataPartition` z pakietu `caret` na uczący i testowy w proporcji 70% i 30%.
- Zachowaj tylko właściwe zmienne i skonwertuj do innego typu danych jeżeli jest taka potrzeba.
- Podzielony i przygotowane zbiory danych zapisz jako `cases_train` i `cases_test`
- Sprawdź jak wygląda rozkład zmiennej celu `IfPayment6M` w zbiorze uczącym i testowym.

Zadanie 3

Utwórz drzewo klasyfikacyjne do modelowania zjawiska dokonania wpłaty w pierwszych 6 miesięcy obsługi. Skorzystaj z przygotowanych danych z zadania 2. Zadanie wykonaj wykorzystując pakiet `rpart`. - Ile węzłów zawiera wygenerowane drzewo? - Wyświetl podsumowanie dla zbudowanego modelu. - Wygeneruj wizualzację drzewa przy użyciu pakietu `rpart.plot`.

Zadanie 4

Zmodyfikuj drzewo klasyfikacyjne z poprzedniego zadania zmieniając wartości parametrów `cp`, `maxdepth`, `minsplit`. Co można zauważyć?

Zadanie 5

Dokonaj predykcji klasy na podstawie zbudowanych modeli drzew zarówno dla zbioru uczącego jak i testowego Wyniki zapisz do zmiennych

Zadanie 6

Wygeneruj macierz konfuzji przy użyciu funkcji `confusionMatrix` z pakietu `caret` dla najbardziej i najmniej złożonego drzewa z poprzedniego zadania dla zbioru uczącego i testowego. Jak kształtują się miary Accuracy, Precision i Sensitivity w obydwu macierzach?

Zadanie 7

Wygeneruj wykres ROC dla zbioru uczącego i testowego za pomocą napisanej przez Ciebie funkcji z zadania 1.

Zadanie 8

Zbuduj drzewo regresyjne w oparciu o dane aplikacyjne i behawioralne z pierwszych trzech miesięcy i oszacuj skuteność skumulowaną od 4 do 12 miesiąca obsługi.

Zadanie 9

Zbuduj drzewa regresyjne do modelowania tego samego problemu jak w zadaniu 7. Tym razem modeluj dla wszystkich kombinacji wartości parametrów `maxdepth = seq(2,6,1)`, `minsplit = c(500,1000,5000)`

Do zbudowania siatki kombinacji wartości parametrów użyj funkcji `expand.grid`.

Zadanie 10

Zbuduj po jednym modelu klasyfikacyjnym i regresyjnym metodą lasów losowych (pakiet `randomForest`). Czy wyniki uzyskane tą metodą są lepsze od najlepszych uzyskanych w zadaniach 6 i 9 (stosując odpowiednio kryterium RMSE i AUC)