[Warsztaty Badawcze] Projekt 1

Mateusz Mechelewski 262760

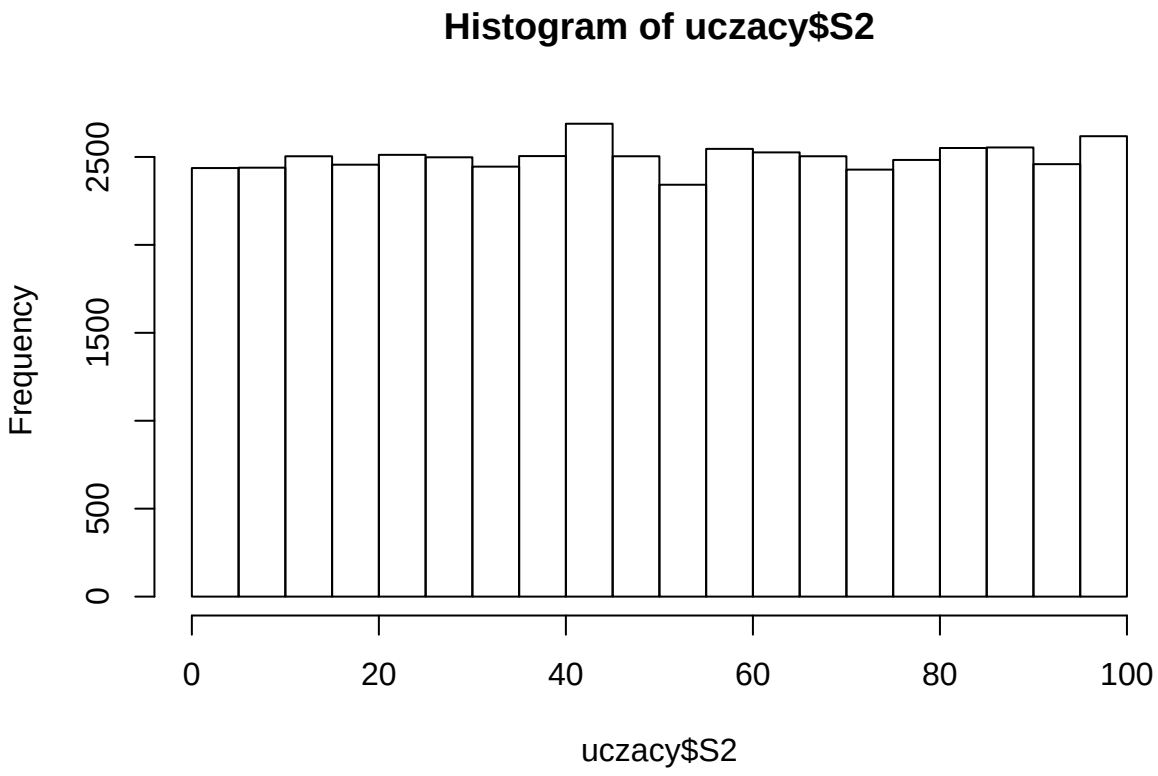
Opis danych

Projekt ma na celu klasyfikację testowego zbioru danych według modelu statystycznego zbudowanego na podstawie zbioru treningowego. Dane zawierają 50 zmiennych oraz zmienną wynikową. Dane testowe należy odpowiednio zakwalifikować do jednej z dwóch klas (`klasa +` lub `klasa -`).

Rozkład zmiennych

Pierwszym krokiem analizy danych było sprawdzenie rozkładu poszczególnych zmiennych. Na podstawie wyniku funkcji `summary` można zauważyć, że dane składają się ze zmiennych jakościowych oraz zmiennych ilościowych (o dwóch lub czterech poziomach). Zmienne jakościowe przyjmują wartości z zakresu [1,100]. Należy jeszcze zwrócić uwagę, że klasy są zrównoważone (dla obu występuje podobna liczba rekordów).

Należy zauważyć, że wartości przyjmowane przez poszczególne zmienne w zbiorze testowym są zrównoważone, o czym świadczy przykładowy histogram przedstawiony dla kolumny `S2`:



Kolejnym krokiem była weryfikacja zależności pomiędzy poszczególnymi zmiennymi. W tym przypadku zastosowanie korelacji nie ma sensu, dlatego potrzebujemy innego narzędzia do sprawdzenia zależności pomiędzy zmiennymi. Ze względu na występowanie zarówno zmiennych ilościowych, jak i jakościowych, użyty został test `chi-kwadrat`. Dla zmiennych ilościowych możliwe również było użycie funkcji `table`.

Selekcja zmiennych

Ze względu na dużą liczbę zmiennych, należało dokonać selekcji zmiennych istotnych. Dokonana ona została na podstawie funkcji `step`. Jako wskaźnik dopasowania modelu wybrane zostało kryterium BIC.

```
model <- glm(y~., data = uczacy, family='binomial')
selekcja <- step(model, k=log(nrow(uczacy)))
```

Jako model docelowy przyjęto formułę `y ~ E1 + M1 + Q1 + W1 + P2 + T2 + U2`.

Miara oceny modeli

Jakość modelu oceniona zostanie na podstawie wyboru 20% rekordów o najwyższym prawdopodobieństwa przynależności do klasy +. Dla wierszy wybranych w ten sposób, obliczona zostanie liczba faktycznie przynależących do klasy dodatniej.

W celu wyboru najlepszego modelu statystycznego dla analizowanego zbioru danych konieczna jest ocena poszczególnych modeli na zbiorze treningowym. Z 50 tysięcy rekordów ze zbioru treningowego wybranych zostało 10 000 losowych wierszy, które stały się nowym zbiorem testowym. Na tej podstawie określona została skuteczność dopasowania każdego z badanych modeli. Poniżej przedstawiony został kod pozwalający na ocenę danego modelu:

```
trainData <- uczacy[1:40000,]
testData <- uczacy[40001:50000,]
modelScore <- getModelScore(formula, trainData, testData)
top <- head(order(modelScore, decreasing = TRUE), 0.2*nrow(testData))
mean(testData[top, "y"])
```

Regresja liniowa

Pierwszym z testowanych modeli statystycznych był model regresji liniowej. Model zbudowany został przy pomocy następującego kodu:

```
getGlmScore <- function(formula, trainData, testData) {
  model <- glm(formula, data=trainData)
  response <- predict(model, testData)
  return(response)
}
```

Wadą tego modelu jest brak odporności na liniową zależność zmiennych. Jakość tego modelu na podstawie miary oceny modelu została oceniona jako 0.727.

Lasy losowe

Drugi z testowanych modeli statystycznych

```
getRandomForestScore <- function(formula, trainData, testData) {
  model <- randomForest(formula, data=trainData, ntree=100)
  response <- predict(model, testData)
  return(response)
}
```

Jakość tego modelu na podstawie miary oceny modelu została oceniona jako 0.8095.

Podsumowanie

Na podstawie powyższych rozważań jako model statystyczny wybrane zostały lasy losowe. Na podstawie tego modelu wyznaczony został podział na klasy dla zbioru testowego. Klasyfikacja ta została przedstawiona w plku `Mateusz_Mechelewski.txt`.