# Warsztaty badawcze - projekt 1

*Magda Tatarynowicz*

## Opis zadania

Zadanie polega na zbudowaniu klasyfikatora dla przykładowych danych, w których każda obserwacja należy do jednej z dwóch klas. Mamy dostępne dwa zbiory - treningowy i testowy, przy czym klasy zbioru testowego nie są znane. Należy zatem dokonać predykcji prawdopodobieństwa, z jakim dane ze zbioru testowego należą do klasy oznaczonej symbolem '+'.

W celu porównania różnych klasyfikatorów dokonałam podziału zbioru treningowego na dwa podzbiory - jeden złożony z 75% obserwacji, który zostanie wykorzystany do trenowania, oraz drugi złożony z pozostałych obserwacji, dla którego dokonamy predykcji oraz sprawdzimy jak mają się wyliczone prawdopodobieństwa do faktycznych wartości klas.

## Selekcja zmiennych

Z racji dużej ilości zmiennych znajdujących się w modelu, najpierw dokonamy ich selekcji. Przeprowadzona zostanie selekcja przy użyciu algorytmu lasów losowych oraz modelu liniowego (funkcja glm).

W przypadku algorytmu lasów losowych wystarczy po dopasowaniu modelu dla zbioru początkowego przeanalizować wykres istotności zmiennych (funkcja varImpPlot). Widać z niego, że warto skorzystać z dziewięciu zmiennych o największej istotności.

Dla modelu liniowego, po dopasowaniu modelu można skorzystać z wartości testu t, który mówi o tym jak istotna jest dana zmienna. W przypadku wydruku dla funkcji summary najbardziej istotne zmienne zostały oznaczone gwiazdkami. Możemy przyjąć, że za istotnie uznajemy zmienne, dla których p-wartość testu jest mniejsza niż 0.05. Na wydruku zostało przedstawionych siedemnaście zmiennych spełniających ten warunek. Warto dodać, że zmienne jakościowe zostały tutaj przerobione na kilka kolumn, wobec czego jeśli porównamy zmienne do tych z modelu oryginalnego, to bierzemy pod uwagę mniej zmiennych.

Dla tak ograniczonych modeli możemy teraz dokonać faktycznej klasyfikacji dla zbioru treningowego.

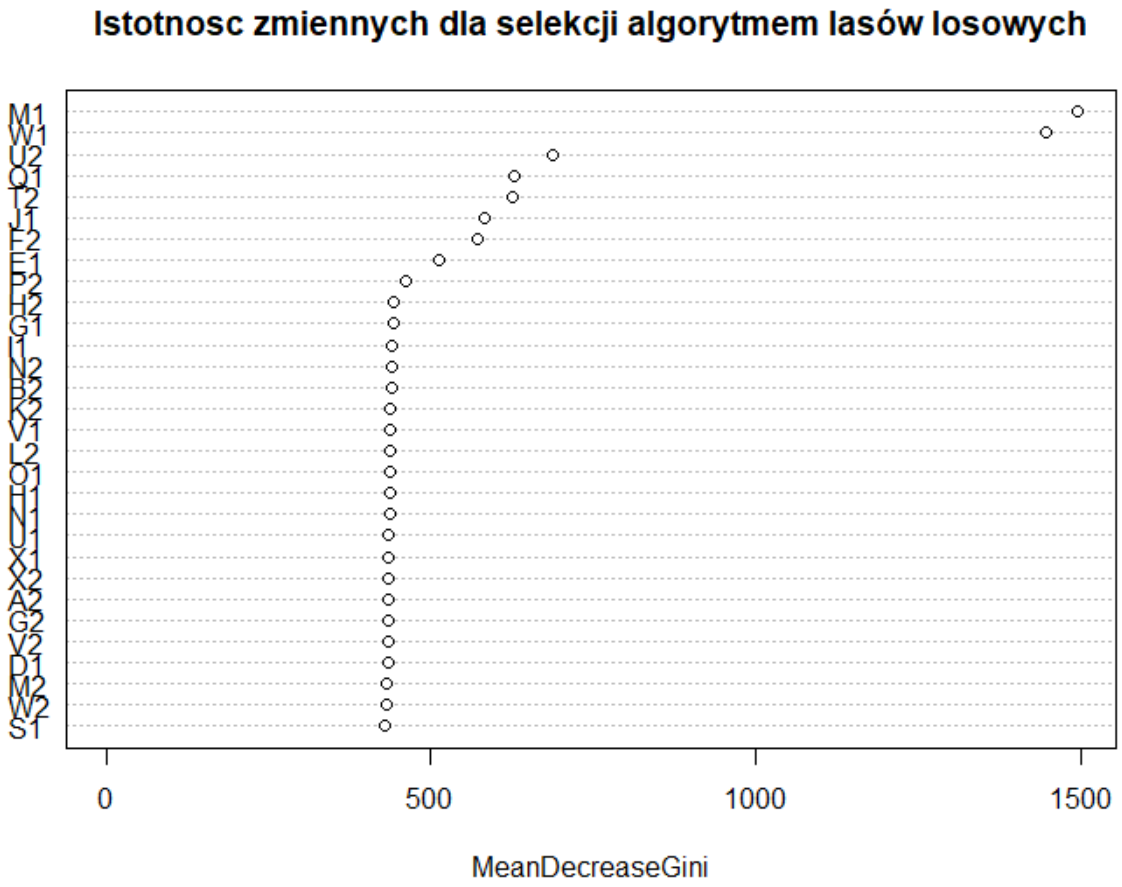


Figure 1: Wykres istnotności zmiennych po wykonaniu algorytmu lasów losowych

## Dopasowanie modeli

Modele klasyfikacyjne zostaną dopasowane dla dwóch algorytmów: lasów losowych oraz xgboost. Dokonamy klasyfikacji dla modeli z ograniczoną liczbą zmiennych oraz dla modelu pełnego w celu porównania jak bardzo różnią się wyniki.

Poniższa tabela przedstawia wyniki dopasowania każdego klasyfikatora, przy czym w celu obliczenia jakości dopasowania wybrano 20% obserwacji ze zbioru testowego z najwyższym obliczonym prawdopodobieństwiem posiadania 'klasy +', a następnie wybrano z nich procent obserwacji, które faktycznie posiadały 'klasę +' (jako że prawdziwe wartości były tu znane).

| Metoda selekcji | Wynik dla xgboost [%] | Wynik dla randomForest [%] |
|---|---|---|
| wszystkie zmienne | 82.92 | 83.48 |
| glm | 81.92 | 80.08 |
| randomForest | 83.72 | 82.76 |

## Analiza wyników i wybór ostatecznego modelu

Z przedstawionych wyników widzimy, iż najlepszą wartość dopasowania udało się uzyskać dla modelu, na którym dokonano selekcji zmiennych metodą lasów losowych, a następnie dokonano klasyfikacji algorytmem xgboost. Warto zauważyć, że różnice dla tej selekcji zmiennych i dwóch wykonanych algorytmów nie są znaczne.

Większą różnicę możemy dostrzec przy porównaniu dwóch metod selekcji zmiennych - w obu przypadkach gorzej sprawdził się model ograniczony funkcją glm, jako że model ten uwzględnia jedynie liniowe zależności między zmiennymi.

Widać także, że dla pełnego modelu nie uzyskaliśmy wcale dużo lepszych wyników, a wręcz dla algorytmu xgboost były one nieco gorsze.

W związku z powyższym jako ostateczny model wybrano model oparty na selekcji zmiennych metodą lasów losowych a następnie sklasyfikowany algorytmem xgboost. Przy użyciu tej metody dopasowano cały dostępny zbiór treningowy, a następnie wykorzystano otrzymany klasyfikator do uzyskania prawdopodobieństw dla zbioru testowego.