Projekt 2 DATA MINING

Marika Partyka

09.06.2019

Po wczytaniu danych treningowych, podzieliłam je na kolejny zbiór treningowy (80%) i testowy (20%).

1 Selekcja metodą CMIM

Na początku testowałam metodę CMIM dla różnych modeli. Pierwszym pomysłem było sprawdzenie modelu logistycznego.

Maksymalna wartość AUC wyniosła około 60% dla wyboru 9 zmiennych.

W ten sam sposób przetestowałam metodę lda, ale ona także nie dała zadowalających wyników: dla 9 kolumn - około 61%.

Następnie sprawdziłam działanie pojedynczego drzewa decyzyjnego dla ustawień cp=0.01 oraz minsplit = 5. Wyniki były dużo lepsze, AUC było równe 81% dla 14 zmiennych.

Kolejne 3 metody, których użyłam dawały już bardzo satysfkacjonujące wyniki w porównaniu do poprzednich metod. Pierwszą metodą był bagging z pakietu adabag.

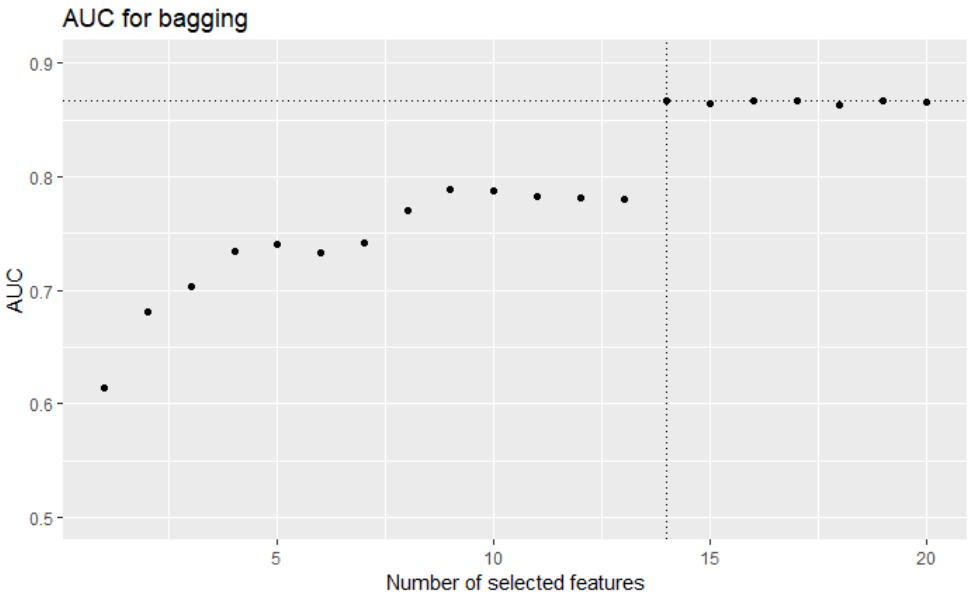
Wykres 1 przedstawia wartość wskaźnika AUC w zależności od liczby wyselekcjonowanych zmiennych. Największe AUC = 0.87 mamy dla 14 zmiennych.

Następny algorytm to AdaBoost z tego samego pakietu. Dla 16 kolumn wskaźnik AUC wynosił 92% - wykres 2.

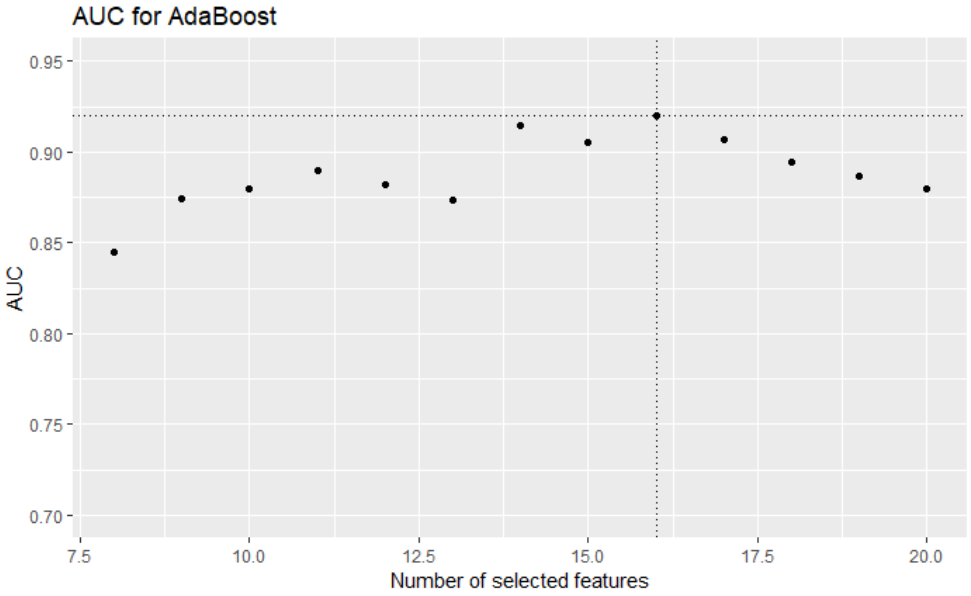
Ostatnim algorytmem, który testowałam dla wyboru zmiennych metodą CMIM był las losowy przy ustawieniach domyślnych. Dla 16 zmiennych wynik był zadowalający: 94% - wykres 3.

2 Selekcja metodą Boruta

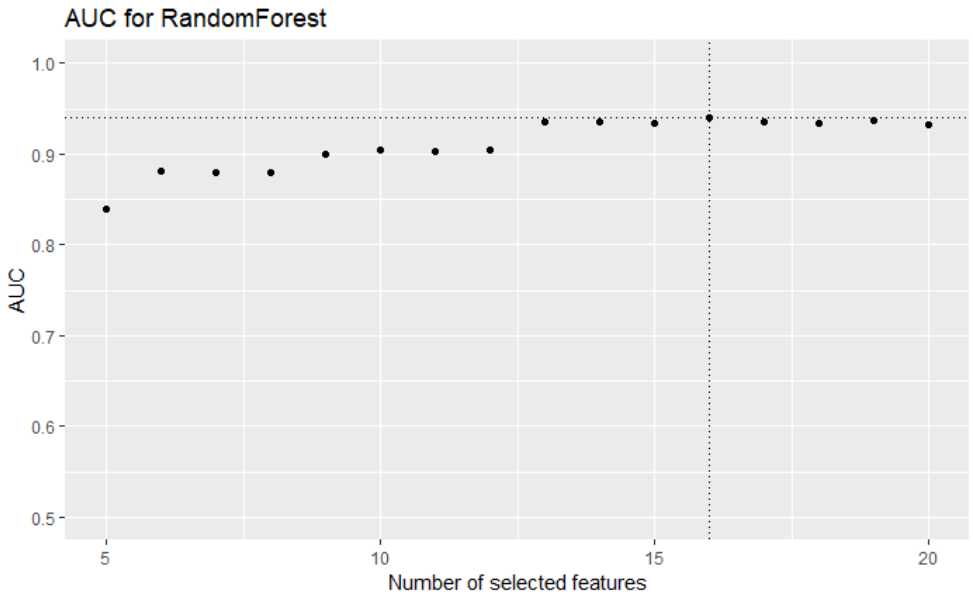
Kolejną metodą selekcji zmiennych była metoda Boruta. Wybór 20 zmiennych. Na nowym zbiorze przetestowałam 3 metody: RandomForest, Bagging i AdaBoost. Wykres 4 przedstawia średnie AUC przy 10 próbach dopasowania każdego z 3 modeli. Średnio, najlepiej sprawdzała się metoda lasu losowego i tę metodę przyjęłam za ostateczną.



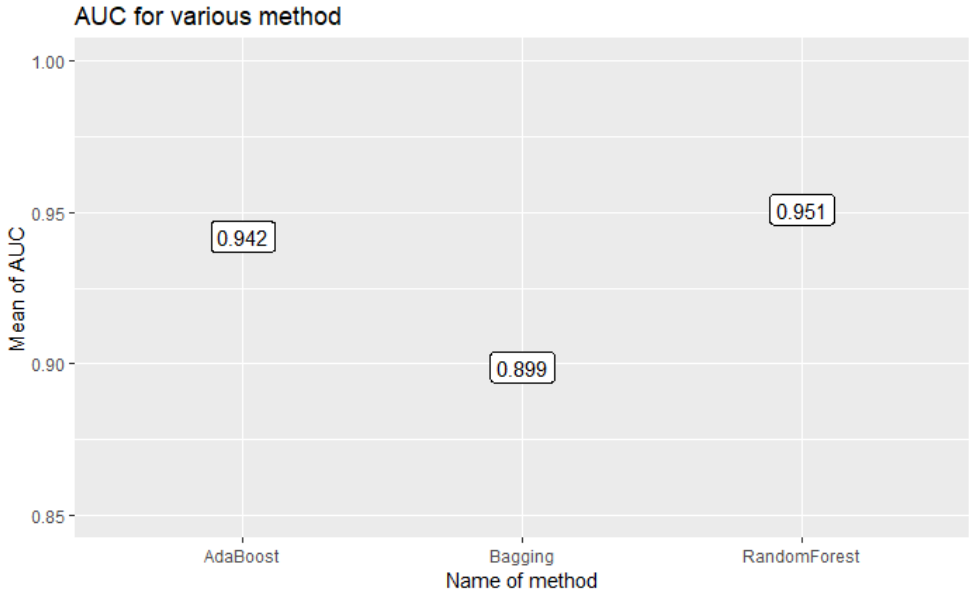
Rysunek 1: Wykres 1



Rysunek 2: Wykres 2



Rysunek 3: Wykres 3



Rysunek 4: Wykres 4