Projekt 1

Magda Tatarynowicz i Kamil Romaszko

Wynik

Zadanie polega na zbudowaniu klasyfikatora dla przykładowych danych, w których każda obserwacja należy do jednej z dwóch klas. Należy dokonać predykcji prawdopodobieństwa, z jakim dane ze zbioru testowego należą do klasy oznaczonej symbolem '+'.

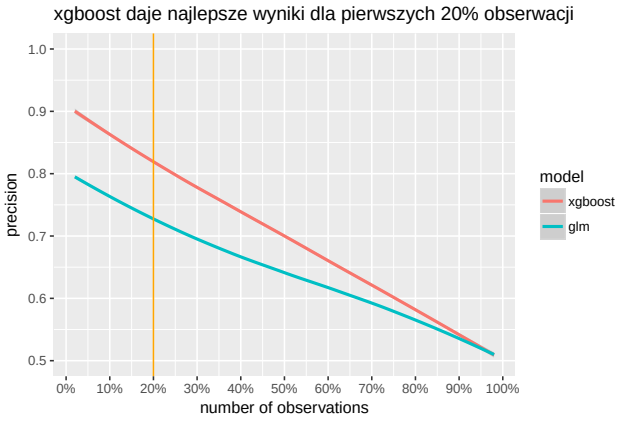
Aby przetestować wybrane metody, dane podzielono na zbiór treningowy i testowy w stosunku 4:1.

Każdy model wyznacza prawdopodobieństwo przynależności do klasy "+". Do oceny modelu, dokonujemy predykcji na zbiorze testowym, a następnie wybieramy 20% najbardziej prawdopodobnych obserwacji i mierzymy procent obserwacji, które faktycznie należą do klasy "+". Im więcej poprawnie przydzielonych obserwacji tym lepiej.

Dla analizowanych modeli, najlepszy wynik uzyskał model **xgboost** z poprawnością **85.2%**.

Poprawność dopasowania

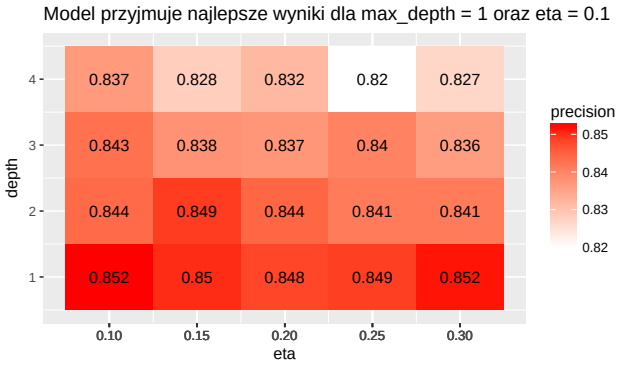
Porównaliśmy też jakość dopasowania dla modelu glm oraz xgboost. Wynik zaprezentowany jest na poniższym wykresie:



Lista rozważanych modeli

W rozważanym modelu, którego wynik zaprezentowano na początku, najpierw dokonano selekcji zmiennych przy użyciu algorytmu lasów losowych. Na podstawie jego wyniku możemy stwierdzić, że tylko kilka pierwszych zmiennych jest istotna. Do dalszych rozważań wybrano 10 najbardziej istotnych zmiennych.

Sam klasyfikator został zbudowany przy użyciu algorytmu xgboost z parametrem *nround=500* co daje dostatecznie złożony model, a jednocześnie sprawia, że wyliczany jest on stosunkowo szybko. Metoda przyjmuje także opcjonalnie parametry *max_depth* oraz *eta* od których zależy dokładność modelu:



Uzasadnienie poprawności

Jak widać, wynik uzyskany przez model jest stosunkowo wysoki, co wynika z przyjętych metod.

Algorytm lasów losowych użyty do selekcji zmienncyh poprawnie wybrał najważniejsze zmienne, wykrywając jednocześnie nieliniowe zależności pomiędzy zmiennymi.

Algorytm xgboost, który zastosowano do budowy modelu został z kolei wykonany dla różnych wartości parametrów *max_depth* i eta i dla nich wybrano najlepsze dopasowanie modelu.