Projekt 1

Anton Lenartovich, Mateusz Mechelewski

Uzyskany wynik

Jako najlepszy model wybrany został **xgboost** z wyborem zmiennych według kryterium BIC - uzyskano dla niego skuteczność na poziomie **88%**.

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.

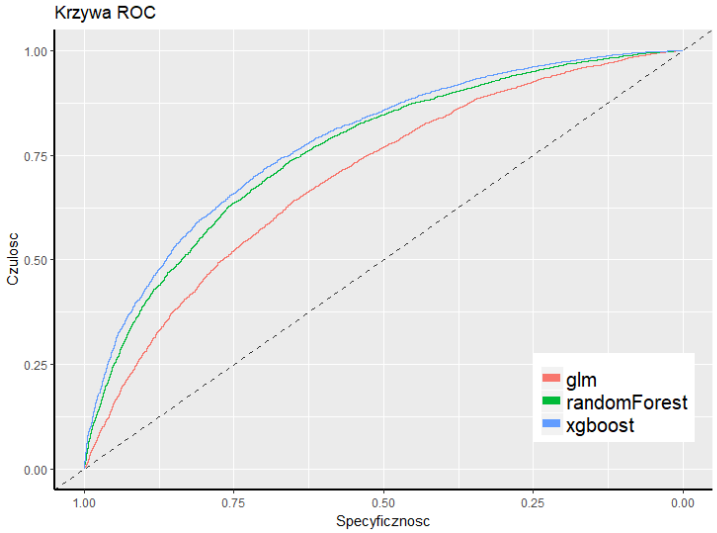
W celu wyboru najlepszego modelu statystycznego dla analizowanego zbioru danych konieczna jest ocena poszczególnych modeli na zbiorze treningowym. Z 50 tysięcy rekordów ze zbioru treningowego wybranych zostało 10 000 losowych wierszy, które stały się nowym zbiorem testowym.

Selekcja zmiennych

Ze względu na dużą liczbę zmiennych, należało dokonać selekcji zmiennych istotnych. Przyjęte zostały dwa podejścia. W pierwszym jako wskaźnik dopasowania modelu wybrane zostało kryterium BIC - uzyskano formułę `y ~ E1 + M1 + Q1 + W1 + P2 + T2 + U2`. Innym zestawem zmiennych jest `y ~ M1 + W1 + U2 + T2 + Q1 + F2 + J1 + E1` wybrany przez metodę bazującą na lasach losowych.

Porównanie klasyfikatorów

Porównując otrzymane wyniki widzimy, że najlepiej poradził sobie model xgBoost. Funkcja oceniająca poprawność klasyfikacji sprawdzała wynik dla 20% danych o największym współczynniku score. Warto również zauważyć, że w większości przypadków lepsze wyniki były otrzymywane dla modeli o zmiennych wybranych za pomocą kryterium BIC.



	rf	BIC
glm	0.6965	0.7294
randomForest	0.8541	0.8447
xgBoost	0.8376	0.88