WB01

Łukasz Ławniczak

16 października 2017

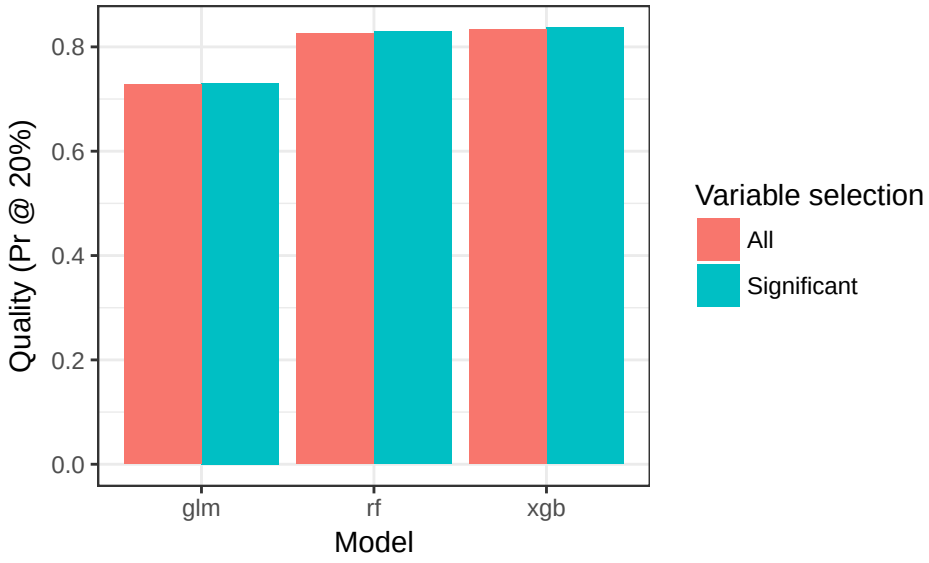
Wynik

Rozważane zadanie polegało na **klasyfikacji binarnej** sztucznie wygenerowanego zbioru danych ze **zrównoważonymi klasami**. Model powinien przewidywać **prawdopodobieństwo**, że dana próbka należy do *klasy +* na podstawie zbioru atrybutów o nazwach *A1,..,Z1,A2,..,X2*. W zbiorze występują zmienne jakościowe (20) i ilościowe (30).

Najlepszy okazał się model stworzony przez algorytm **Xgboost**, uzyskując skuteczność klasyfikacji na poziomie **83.7%**.

Lista rozważanych modeli / zmiennych

Rozważono następujące rodzaje modeli: *regresja logistyczna*, *lasy losowe* i *gradient boosting*. Modele zostały wygenerowane na **wszystkich zmiennych**, jak również na podzbiorze **zmiennych istotnych**. Istotne zmienne zostały wyznaczone przy pomocy istotności w **lesie losowym** wygenerowanym na wszystkich zmiennych. Modele utworzone z wykorzystaniem jedynie istotnych zmiennych charakteryzują się podobną jakością do modeli opartych na wszystkich zmiennych.



Argumentacja poprawności

Wykorzystanym wskaźnikiem jakości klasyfikatora jest **precyzja** dla **20%** próbek o najwyższych prawdopodobieństwach zwróconych przez model. W celu jego obliczenia zastosowano **walidację prostą**, dzieląc zbiór danych w proporcji **1:1**. Przy doborze parametrów modelu sprawdzano również jakość klasyfikatora na zbiorze uczącym w celu uniknięcia przeuczenia modelu.