Executive report - projekt 1

K. Szczawiński, K. Prusinowski, P. Pollak

Najlepszy model

Najlepszym rozwiązaniem jest wykorzystanie xgboost na zmiennych W1, M1, E1, P2, U2, Q1, T2, J1 i F2. Poprawność klasyfikacji dla 20% danych testowych o największym prawdopodobieństwie klasy "+" wynosi **83,56%**.

Sprawdzane modele

Selekcja zmienny

Użyte metody selekcji zmiennych:

- glm (istność zmiennych)
- krokowa metoda AIC
- LASSO
- randomForest
- istnotność zmiennych w modelu xgboost

Najistotniejsze zmienne to W1, M1, E1, P2, U2, Q1, T2, J1 i F2.

Predyktory

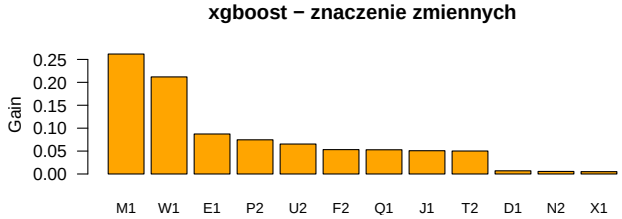
Oceniane algorytmy klasyfikacji to:

- glm
- random forests
- xgboost

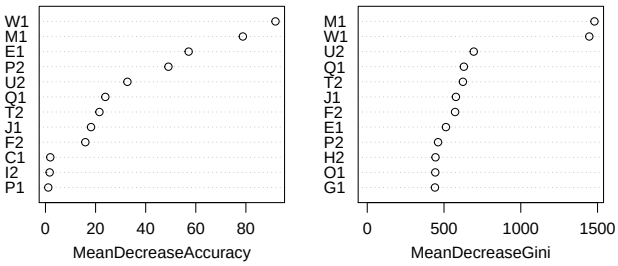
Wyniki poszczególnych modeli

	glm	rf	xgb
Wybrane zmienne	0.7276	0.8268	0.8356
Model pełny	0.7284	0.8304	0.8344

Wizualizacja istotności zmiennych



Wykres przedstawia istostność najważniejszych zmiennych w modelu randomForest.



Argumentacja poprawności

Walidacja

W celu dokonania oceny modeli zbiór treningowy został podzielony na dwie części - treningowy (75% danych) i testowy.

Warto zauważyć, że jakość modelu będzie oceniana na podstawie tego, ile z 20% obserwacji o najwyższym prawdopodobieństwie przynależności do klasy "+" rzeczywiście do tej klasy należy. Nie jest to zwykła dokładność predykcji. Należy to wziąć pod uwagę przy ocenie modeli - są one oceniane właśnie na podstawie 20% danych o najwyższym *score* (prawdopodobieństwie klasy "+", przypisanym przez model).

Wybrane zmienne

Zmienne wybrane przez random forest (pokrywające się z tymi uznanymi przez *xgboost* za najważniejsze) zostaną uznane za najlepiej wyselekcjonowane. W dużej mierze pokrywają się z tymi wybranymi na podstawie różnych metod opartych o *glm*. Jednak *glm* nie jest w stanie wykryć zależności innych niż liniowe. Zatem to zmienne "W1", "M1", "E1", "P2", "U2", "Q1", "T2", "J1" i "F2" zostały przyjęte jako istotne.