Warsztaty badawcze - projekt 1

Paweł Pollak

14 października 2017

Dane

Dane zostały wygenerowane w sposób sztuczny, zatem nie da się interpretować znaczenia różnych kolumn. Zmienna odpowiedzi przyjmuje jedynie dwie wartości, zatem mamy do czynienia z klasyfikacją binarną. Grupy są zbalansowane, obserwacji o przypisanej klasie "+" jest `49.986`%.

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).

Selekcja zmiennych

glm

Zmienne o największej istotności (na podstawie p-wartości odpowiadającej danej zmiennej) to A1, E1, M1, Q1, W1, P2, T2 i U2.

AIC na glm

Selekcja dokonana metodą krokową, zaczynając od modelu pełnego.

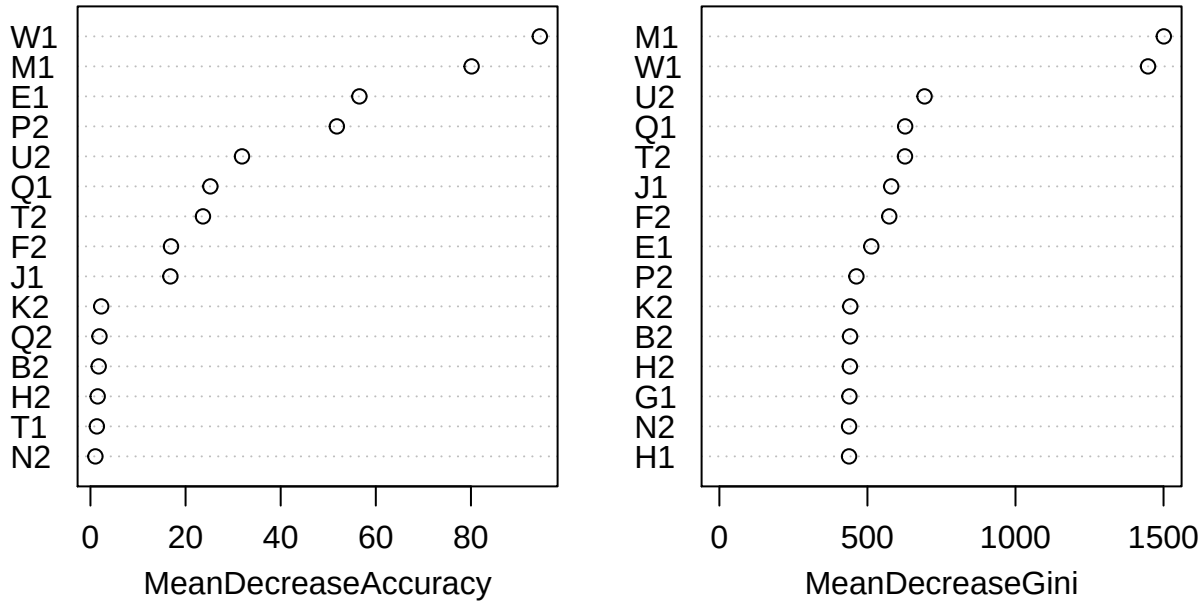
Wybrane przez funkcję zmienne (zaczynając od modelu pełnego) to A1, B1, E1, F1, M1, P1, Q1, W1, Y1, Z1, P2, T2 i U2.

LASSO

Zmienne wybrane wykorzystując LASOO to A1, E1, F1, M1, Q1, W1, Z1, P2, T2 i U2.

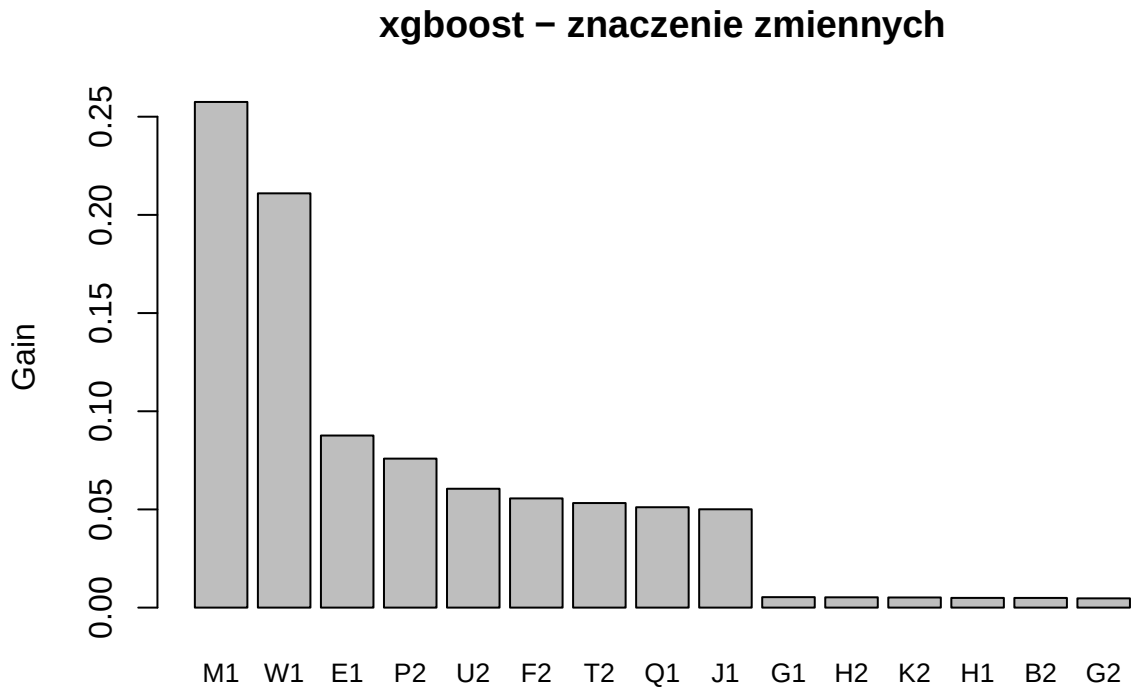
Random forest

Wykres uwzględnia 15 najbardziej znaczących zmiennych (wykres został utworzony za pomocą funkcji *varImpPlot*.



Zmienne, które model uznał za najbardziej znaczące to W1, M1, E1, P2, U2, T2, Q1, F2, J1.

xgboost



Wykres przedstawia znaczenie zmiennych dla modelu budowanego przez xgboost. Te o największym wpływie na predykcję są identyczne jak w przypadku lasów losowych.

Podsumowanie

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" zostaną przyjęte jako istotne.

Predykcja

Oceniane algorytmy klasyfikacji to:

- glm
- random forests
- xgboost

Zostały one poddane ewaluacji na podstawie omówionej wcześniej miary na:

- modelu pełnym (uwzględniającym wszystkie zmienne)
- modelu zawierającym jedynie zmiennej, wybrane w poprzedniej części raportu

Tabela przedstawia wyniki poszczególnych metod na odpowiednich modelach.

	glm	rf	xgb
Wybrane zmienne	0.7276	0.8264	0.8400
Model pełny	0.7284	0.8340	0.8304

Jak widać, selekcja zmiennych okazała się obniżać dokładność klasyfikacji w ramach rankingu dla metod innych niż xgboost. Może to wynikać z faktu, że obserwacje zostały sztucznie uzyskane i bardzo trudno jest wychwycić, które zmienne są rzeczywiście istotne.

Ostatecznie do klasyfikacji zbioru testowego zostanie wykorzystany xgboost uwzględniający jedynie zmienne, które zostały uznane za istotne.