Projekt 1 - Raport

Karol Szczawiński

14 października 2017

1. Selekcja cech

Na wstępie zbiór danych podzielono na dwa podzbiory: treningowy i testowy w stosunku 4:1. Do selekcji zmiennych na zbiorach treningowych użyto dwóch metod.

1.1. Lasy losowe

W związku z tym, że w danych występują zmienne kategoryczne, które przyjmują wartości dyskretne to w przypadku lasów losowych stworzono dwa modele. Jeden na danych niezmienionych, drugi w którym zmienne kategoryczne zostały zamienione na kolumny binarne (one hot encoding). Na ich podstawie wybrano następujące kolumny:

1. M1 + W1 + U2 + Q1 + T2 + F2 + J1 + E1 (model 1)

2. M1 + W1 + U2 + Q1 + T2 + F2 + J1 (model 2)

1.2. Selekcja krokowa dla regresji logistycznej

W przypadku regresji logistycznej stworzono tylko jeden model po przetworzeniu kolumn kategorycznych jak w przypadku lasów losowych. Następnie wybrano z niego najbardziej istotne zmienne przy użyciu selekcji krokowej z karą BIC. Stworzono następujący model danych:

E1B + E1C + E1D + M1 + Q1 + W1 + P2B + P2C + P2D + T2 + U2 (model 3)

2. Użyte algorytmy

- Naiwny algorytm Bayesa
- Lasy losowe
- XGBoost

3. Rezultaty

Dla każdego z modeli danych użyty został każdy algorytm (dla XGBoost nie można było tylko użyć danych z kolumnami kategorycznymi). Ocenę predykcji dokonano na podstawie precyzji dla 20% rekordów o największym prawdopodobieństwie. Końcowe wyniki podano tylko dla parametrów, które uzyskane najlepszą ocenę. Poniżej dla każdego z algorytmów otrzymane wyniki.

	Model 1	Model 2	Model 3
Naiwny algorytm Bayesa	0.781	0.749	0.789
Lasy losowe	0.821	0.796	0.821
XGBoost	0	0.81	0.843

4. Podsumowanie

Na podstawie uzyskanych wyników wybrano model danych stworzony przy użyciu selekcji krokowej oraz algorytm XGBoost. Ponieważ dla nich uzyskano najlepszy wynik i Użyto takiego modelu do predykcji na danych testowych.