Zaawansowane metody uczenia maszynowego
Projekt 2

Katarzyna Spalińska

07.06.2019

1 Cel projektu

Celem projektu jest przetestowanie różnych metod selekcji zmiennych w problemie klasyfikacji.

2 Wykorzystane narzędzia

Do zrealizowania projektu wykorzystano język programowania Python wraz z bibliotekami:

- pandas (obróbka zbiorów danych)
- numpy (obróbka zbiorów danych)
- sklearn (algorytmy uczenia maszynowego oraz selekcji zmiennych)

3 Wstępne przetwarzanie danych

Początkowe wymiary zbioru treningowego wynosiły 501 kolumn x 2 tys. wierszy, jednak ponieważ kolumna z indeksem 500 składała się wyłącznie z brakujących wartości, została od razu usunięta, co dało rzeczywiste początkowe wymiary: 500 kolumn x 2 tys. wierszy. Ponieważ zbiór nie zawierał braków danych ani danych kategorycznych, dalsza część przygotowania danych jest już ściśle związana z selekcją zmiennych.

4 Selekcja zmiennych

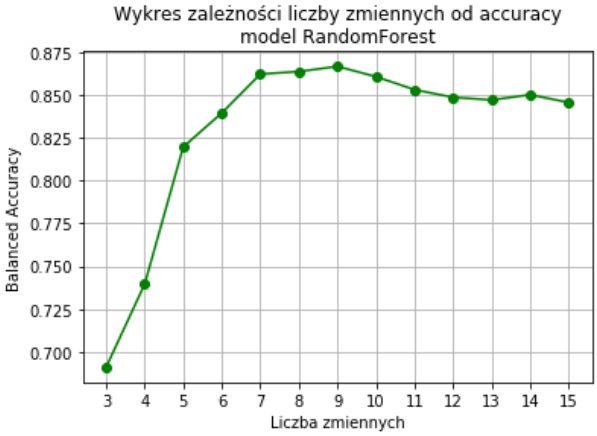
Pierwszym etapem selekcji zmiennych było wybranie i usunięcie kolumn skorelowanych ze sobą. Sprawdzono korelacje pomiędzy wszystkimi kolumnami i jeśli korelacja pomiędzy wybraną parą była większa niż 0.95, wtedy jedna z kolumn zostawała usunięta.

Następnie przetestowano dwie metody selekcji zmiennych, wykorzystując funkcje oferowane przez pakiet *sklearn*.

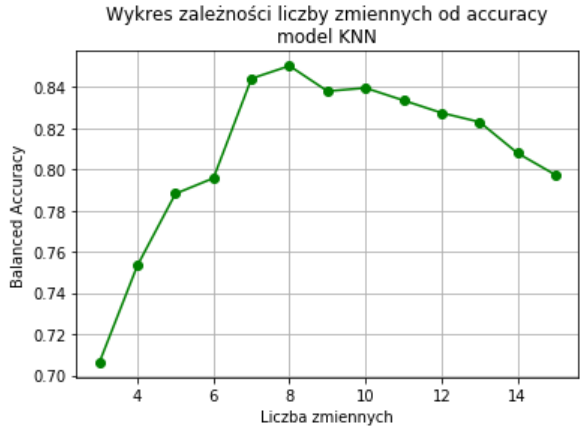
4.1 Metoda KBest

Biblioteka scikit-learn zapewnia klasę SelectKBest, która może być używana z zestawem różnych testów statystycznych, aby wybrać określoną liczbę zmiennych. Wykorzystano test statystyczny chi kwadrat dla cech nieujemnych w celu wybrania n najlepszych cech ze zbioru danych.

Poniżej przedstawiono wykresy zależności skuteczności dwóch przetestowanych modeli (Random Forest, KNN) od liczby zmiennych.



Rysunek 1: Zależność skuteczności od liczby zmiennych w metodzie KBest - model Random Forest

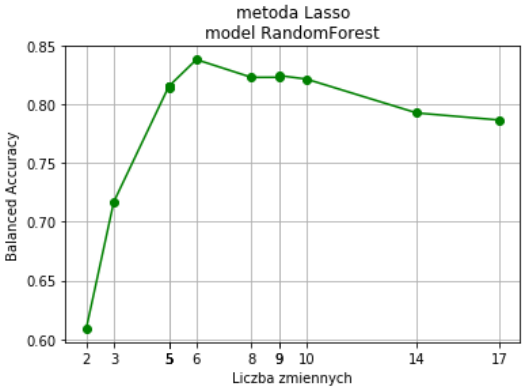


Rysunek 2: Zależność skuteczności od liczby zmiennych w metodzie KBest - model KNN

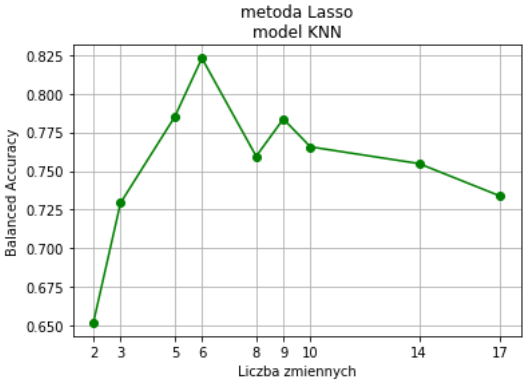
4.2 Metoda Lasso

Wykorzystano zaimplementowaną w bibliotece sklearn metodę Lasso, wybierającą najlepsze zmienne do modelu.

Poniżej przedstawiono wykresy zależności skuteczności dwóch przetestowanych modeli (Random Forest, KNN) od liczby zmiennych.



Rysunek 3: Zależność skuteczności od liczby zmiennych w metodzie KBest - model Random Forest



Rysunek 4: Zależność skuteczności od liczby zmiennych w metodzie KBest - model KNN

Lepsze wyniki balanced accuracy osiągała metoda KBest, jednak pojawiły się trudności w zreplikowaniu wyników modelu po wywołaniu go drugi raz na wybranych cechach (w celu dodatkowego sprawdzenia). Wyższe wyniki mogły być w takim razie przekłamane wcześniejszym dopasowaniem danych, które występuje w tej metodzie (funkcja fit).

Zdecydowano się na wybranie modelu z 6 zmiennymi wybranymi przez metodę Lasso. Wybrane zmienne to: [48, 64, 105, 241, 338, 453].