Project 1

PiotrKrzeszewski

October 9, 2017

Analiza zmiennych

Celem zadania jest przeprowadzenie klasyfikacji.

W dostarczonym zbiorze danych znajduje się 50 zmiennych oraz klasa danego rekordu. 30 z nich to zmienne ilościowe, a pozostałe są jakościowe (zawierające dwie lub 4 wartości). Wszystkie zmienne mają rozkład jednostajny.

Transformacja zmiennych

Dostarczone dane nie potrzebowały dużej ilości przekształceń. Na zbiorze testowym zamieniono kolumnę `y` zawierającą klasę wynikową na kolumnę `res` typu binarnego (rozpoznając, czy pole zawiera znak `+`).

Selekcja zmiennych

Pierwszym etapem przygotowywania klasyfikatora okazała selekcja zmiennych. W tym celu wykorzystano metodę `BIC` w trybie `forward`. Model bazowy okazał pusty i selekcja wybrała 7 zmiennych.

Wybrane zmienne to: `W1, U2, E1, P2, Q1, T2, M1`.

Zostały one wykorzystane w dalszej analizie.

Klasyfikacja

Regresja logistyczna

Na wybranym zbiorze zmiennych przeprowadzono najpierw klasyfikację przy użyciu regresji logistycznej - metoda `glm()` w R. Wyniki tej metody nie okazały się wystarczająco dobre. Przy podziale danych testowych w stosunku 70/30 na dane testowe i uczące udało się otrzymać skuteczność mierzoną metryką zadania na poziomie (zależy od wyboru części treningowej i testowej):

```
## [1] 0.725
```

XGBoost

Kolejną przetestowaną metodą jest XG-Boost. Jest to popularna metoda klasyfikacji, która w niektórych porównaniach dawała najlepsze wyniki. Opiera się ona na zbudowaniu kilku słabych klasyfikatorów, które docelowo stworzą jeden klasyfikator dający dobre wyniki.

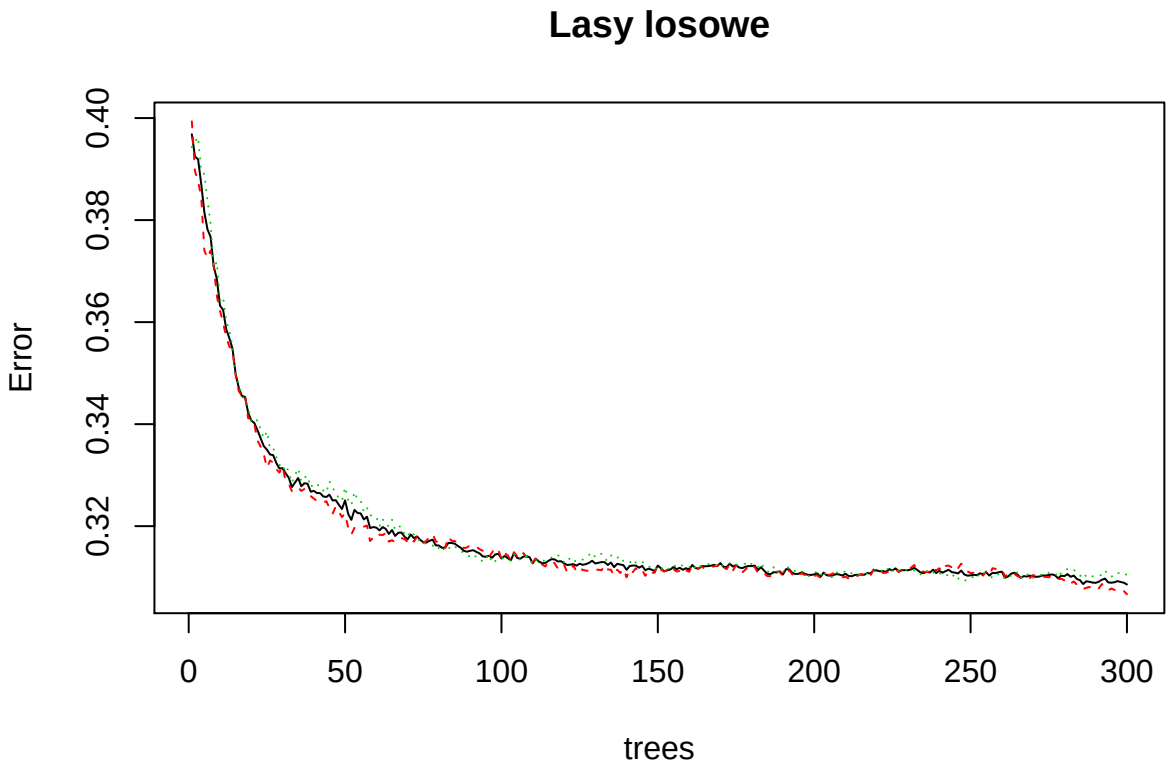
W opisywanej sytuacji XG-Boost wykorzystał dane ze zmiennych, które zostały wybrane podczas selekcji. Otrzyman skuteczność, liczona według warunków zadania to:

```
## [1] 0.8183333
```

Random Forest

Trzecia przetestowana metoda wykorzystuje lasy losowe. Ponownie wykorzystano zmienne wybrane na etapie selekcji. Otrzymana skuteczność klasyfikacji to:

```
## [1] 0.7116667
```



Powyższy wykres przedstawia zależność błędu od ilości zstosowanych drzew. Widać, że znaczne zwiększanie ilości drzew po przekroczeniu 150 nie daje znacznej poprawy dokładności.

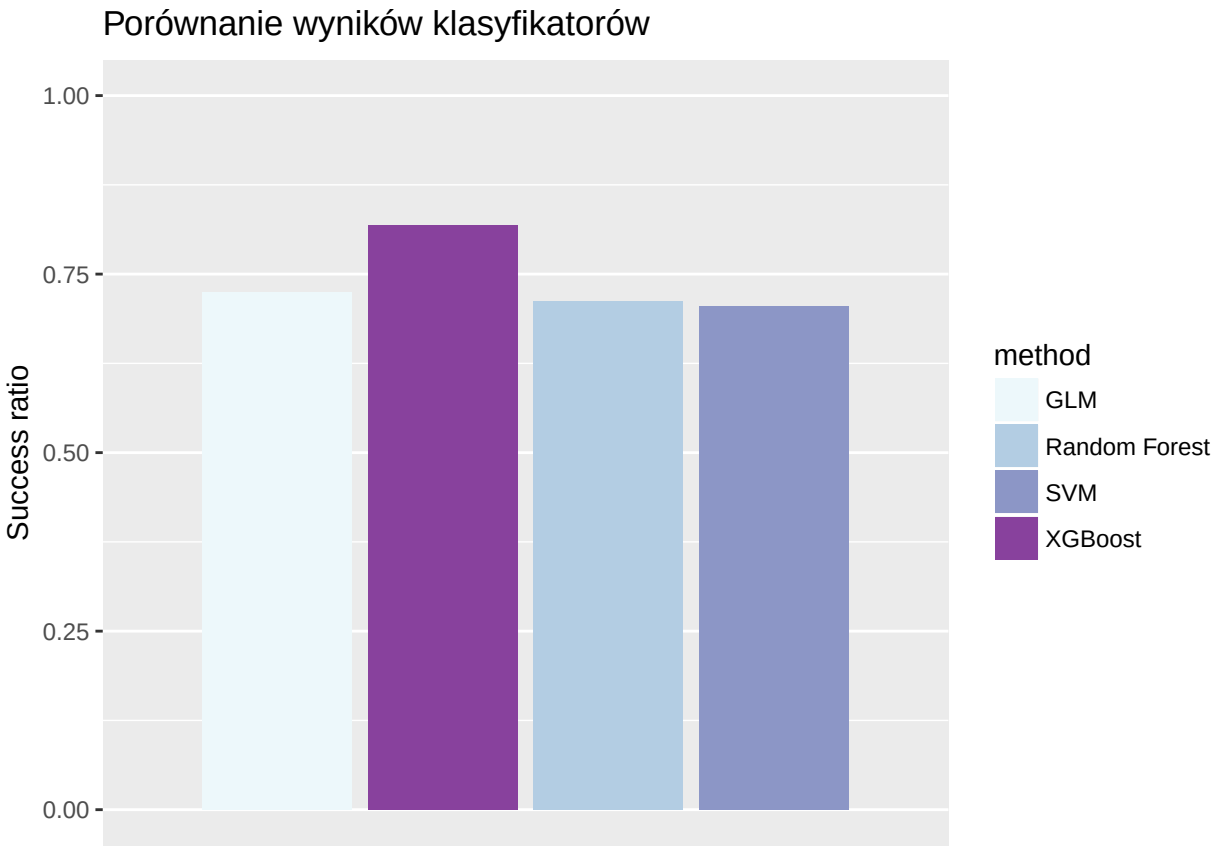
SVM

Ostatnim wykorzystanym modelem jest SVM. Maszyna wektorów nośnych opiera się na podziale n-wymiarowej hiperpłaszczyzny i znalezieniu jej optymalnego podziału

```
## [1] 0.7053333
```

Podsumowanie

Z przeprowadzonych testów wyraźnie wynika, że najlepiej w praktyce sprawdza się algorytm XG-Boost. Zapewnił on ponad 80% skuteczności w przeprowadzonych testach. Dodatkowo wyniki poprawiła selekcja zmiennych przeprowadzona na początku procedury.


Porównanie wyników klasyfikatorów
Success ratio
1.00
0.75
0.50
0.25
0.00
method
GLM
Random Forest
SVM
XGBoost