29 stycznia 2020

Politechnika Wrocławska

Wykrywanie oszustw na kartach płatniczych z wykorzystaniem metod wrażliwych na koszt

Promotor: dr inż. Andrzej Giniewicz

Wydział Matematyki Politechniki Wrocławskiej

Patryk Wielopolski

Wstęp

Problem detekcji oszustw na kartach kredytowych

Zagadnienie:

- Detekcja nieautoryzowanych transakcji.

Problem detekcji oszustw na kartach kredytowych

Zagadnienie:

- Detekcja nieautoryzowanych transakcji.

Aktualna metodologia:

- Reguły eksperckie,
- Modele predykcyjne.

Problem detekcji oszustw na kartach kredytowych

Zagadnienie:

- Detekcja nieautoryzowanych transakcji.

Aktualna metodologia:

- Reguły eksperckie,
- Modele predykcyjne.

Problem:

- Reguły eksperckie same nie wykrywają nowych zachowań przestępców.
- Standardowe modele predykcyjne nie biorą pod uwagę kosztu popełnienia błędu.

Problem detekcji oszustw na kartach kredytowych

Zagadnienie:

- Detekcja nieautoryzowanych transakcji.

Aktualna metodologia:

- Reguły eksperckie,
- Modele predykcyjne.

Problem:

- Reguły eksperckie same nie wykrywają nowych zachowań przestępców.
- Standardowe modele predykcyjne nie biorą pod uwagę kosztu popełnienia błędu.

Rozwiązanie:

- Modele predykcyjne wrażliwe na koszt.

Część teoretyczna

Miary skuteczności modeli

	Stan sprzyjający $y_i = 1$	Stan niesprzyjający $y_i = 0$
Predykcja pozytywna $c_i = 1$	TP	FP
Predykcja negatywna $c_i = 0$	FN	TN

Macierz pomyłek.

$$\text{Precyzja} = \frac{TP}{TP + FP}$$

$$\text{Czułość} = \frac{TP}{TP + FN}$$

$$F_1 = \left(\frac{2}{\text{Precyzja}^{-1} + \text{Czułość}^{-1}} \right)$$

Miary skuteczności modeli wrażliwych na koszt

	Stan pozytywny $y_i = 1$	Stan negatywny $y_i = 0$
Predykcja pozytywna $c_i = 1$	$C_{1,1}^{(i)}$	$C_{1,0}^{(i)}$
Predykcja negatywna $c_i = 0$	$C_{0,1}^{(i)}$	$C_{0,0}^{(i)}$

Macierz kosztu dla i-tej obserwacji.

Oznaczenia:

- $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ – wektor prawdziwych stanów klasyfikacji,
- $\boldsymbol{c} = (c_1, c_2, \ldots, c_N)$ – wektor przewidywanych klas,
- $\boldsymbol{C} = (C_1, C_2, \ldots, C_N)$ – wektor macierzy kosztu,

Miary skuteczności modeli wrażliwych na koszt

$$\text{Oszczędności}(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{C}) = \frac{\text{Koszt bazowy}(\boldsymbol{y}, \boldsymbol{C}) - \text{TC}(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{C})}{\text{Koszt bazowy}(\boldsymbol{y}, \boldsymbol{C})}$$

Oznaczenia:

- Koszt całkowity$(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{C})$ lub $\text{TC}(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{C}) = \sum_{i=1}^{N} C_{c_i, y_i}^{(i)}$
- Koszt bazowy$(\boldsymbol{y}, \boldsymbol{C}) = \min\{\text{TC}(\boldsymbol{y}, \boldsymbol{c}_0, \boldsymbol{C}), \text{TC}(\boldsymbol{y}, \boldsymbol{c}_1, \boldsymbol{C})\} \neq 0$
- $\boldsymbol{c}_0 = (0, 0, \ldots, 0)$ – N-elementowy wektor predykcji równych 0,
- $\boldsymbol{c}_1 = (1, 1, \ldots, 1)$ – N-elementowy wektor predykcji równych 1.

Modele predykcyjne

- Standardowe modele predykcyjne:
  - Regresja logistyczna
  - Drzewo decyzyjne
  - Las losowy
  - XGBoost
- Klasyfikacja wrażliwa na koszt:
  - Minimalizacja ryzyka bayesowskiego
  - Optymalizacja progu
- Trening wrażliwy na koszt:
  - Regresja logistyczna wrażliwa na koszt
  - Drzewo decyzyjne wrażliwe na koszt

Eksperyment

Zbiór danych

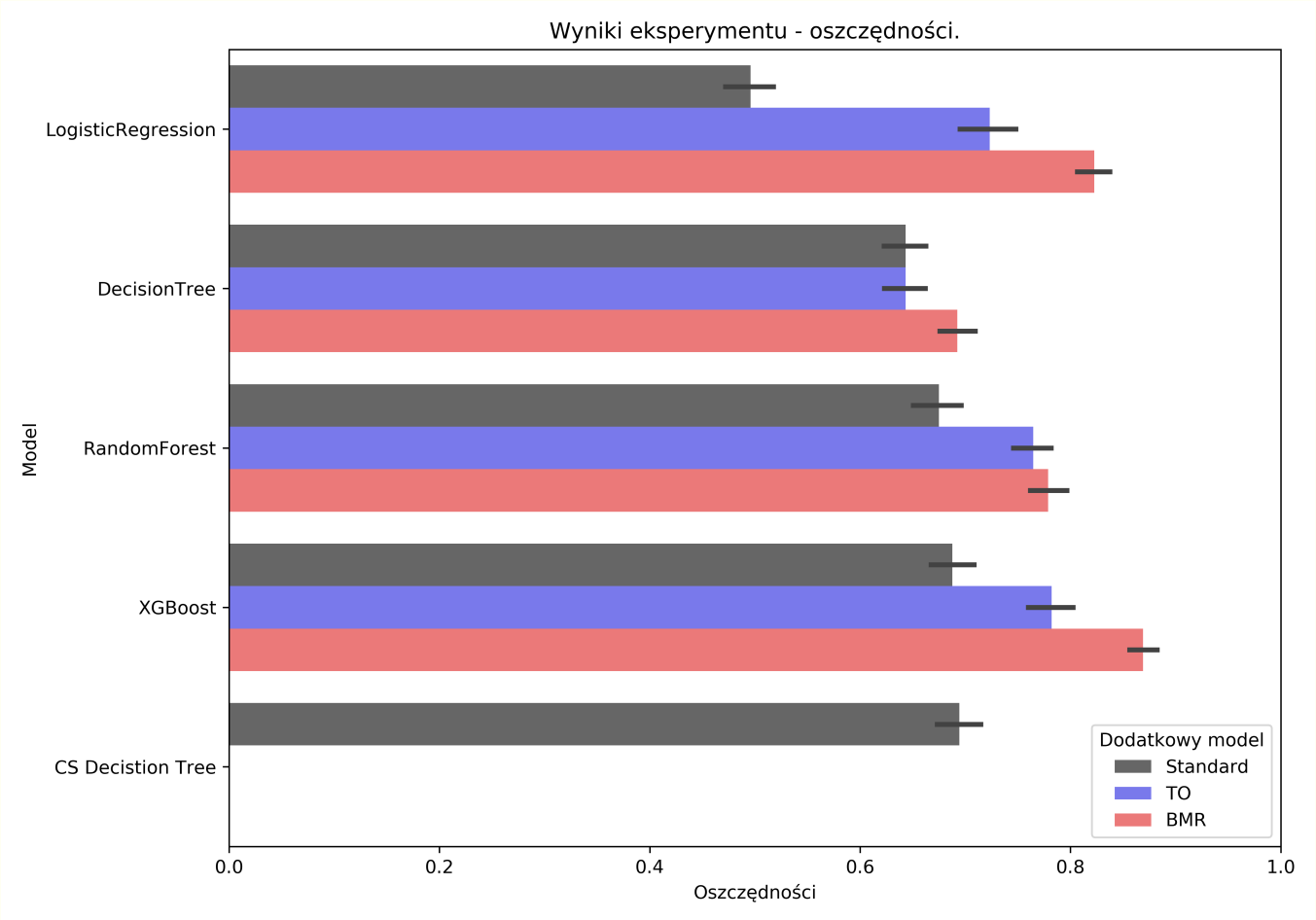
Wykorzystano zbiór danych *Credit Card Fraud Detection*.

- Zawiera transakcje zawarte europejskimi kartami kredytowymi w ciągu dwóch dni we wrześniu 2013 roku.
- Składa się z 284,807 transakcji, w tym z 492 oszustw.
- Obserwacje są opisane 30 atrybutami, w tym 28 z nich to zanonimizowane zmienne numeryczne, które były wcześniej poddane transformacji PCA (*ang. Principal Component Analysis*).

Metodologia eksperymentu

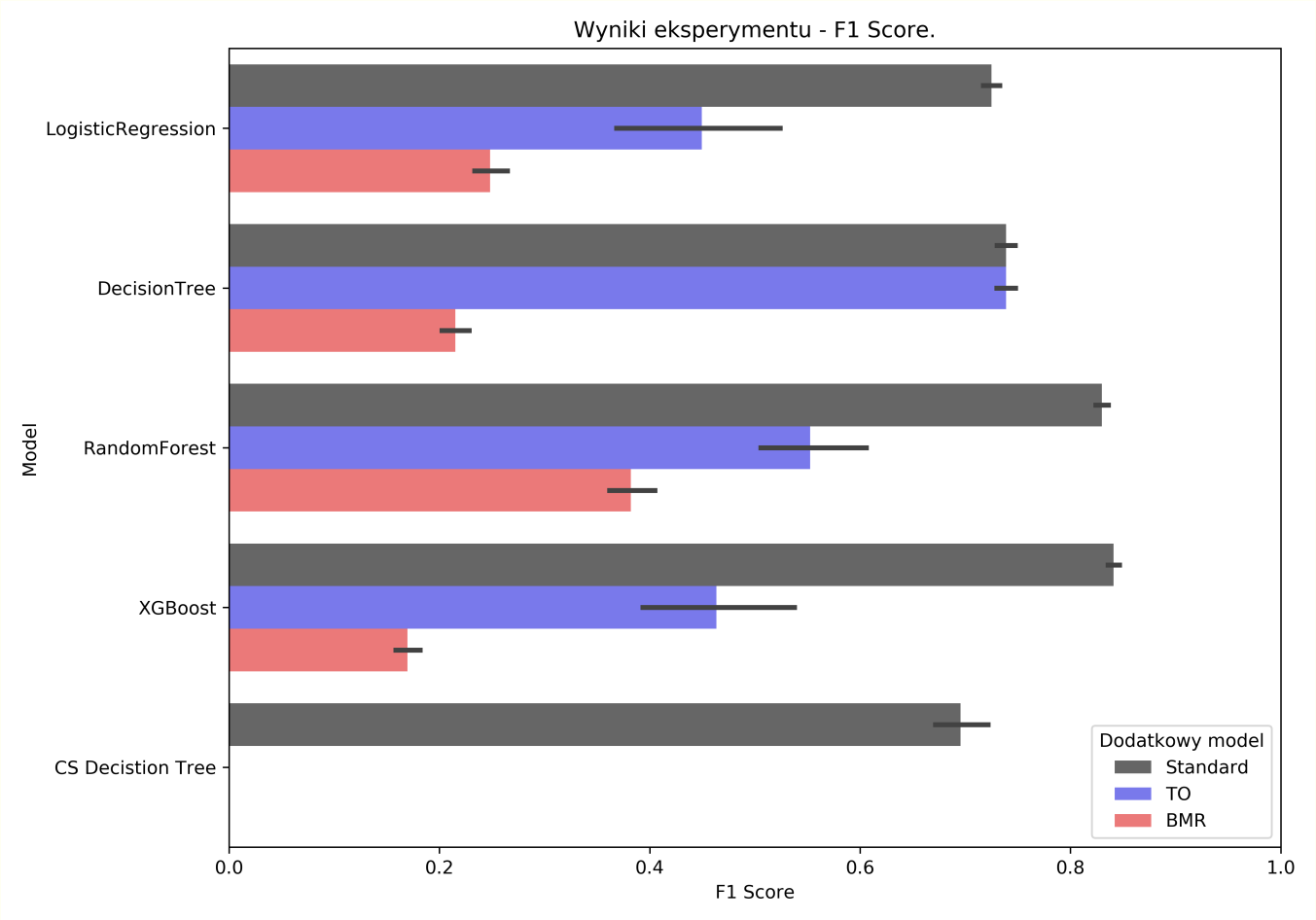
- 50 powtórzeń symulacji Monte Carlo
- Podział zbioru danych:
  - 50% zbiór treningowy
  - 17% zbiór walidacyjny
  - 33% zbiór testowy
- Wykorzystane modele:
  - Modele standardowe: regresja logistyczna, drzewo decyzyjne, las losowy, XGBoost
  - Drzewo decyzyjne wrażliwe na koszt
  - Optymalizacja progu oraz minimalizacja ryzyka bayesowskiego zastosowana dla modeli standardowych

Wyniki dla oszczędności



Wyniki eksperymentu dla oszczędności. Źródło: Opracowanie własne.

Wyniki dla F1 Score



Wyniki eksperymentu dla F1 Score. Źródło: Opracowanie własne.

Wnioski

- Wykorzystanie metod wrażliwych na koszt pozwala na zwiększenie oszczędności.
- Wykorzystanie powyższych metod obniża jakość predykcji w kontekście standardowych miar skuteczności.
- Drzewo decyzyjne wrażliwe na koszt uzyskuje zadowalające wyniki oszczędności, przy jednoczesnym zachowaniu wysokiej jakości predykcji w kontekście standardowych miar skuteczności.

Kierunek dalszych prac

- Analiza metod typu ensemble z drzewem decyzyjnym wrażliwym na koszt jako klasyfikator bazowy.
- Badania oraz implementacja algorytmu XGBoost z drzewem decyzyjnym wrażliwym na koszt jako klasyfikator bazowy.
- Próba wykorzystania miary oszczędności jako niestandardowej funkcji kosztu w algorytmie XGBoost.

Bibliografia


Bahnsen, A. C., Aouada, D., Ottersten, B.
Example-dependent cost-sensitive logistic regression for credit scoring.
In *2014 13th International Conference on Machine Learning and Applications* (Dec 2014), pp. 263–269.


Bahnsen, A. C., Stojanovic, A., Aouada, D., Ottersten, B.
Cost sensitive credit card fraud detection using bayes minimum risk.
In *2013 12th International Conference on Machine Learning and Applications* (Dec 2013), vol. 1, pp. 333–338.


Correa Bahnsen, A., Stojanovic, A., Aouada, D., Ottersten, B.
Improving credit card fraud detection with calibrated probabilities.
In *Proceedings of the 2014 SIAM International Conference on Data Mining. Pennsylvania, USA* (04 2014).

Dziękuję za uwagę!