



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

Możliwości popełnienia przestępstwa:

- ❖ przekazanie numeru karty nieznajomemu,
- ❖ utrata lub kradzież karty,
- ❖ skopiowanie danych karty,
- ❖ kradzież przesyłki z kartą.

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.

$$\text{Koszt całkowity}(\mathbf{y}, \mathbf{c}, \mathbf{C}) = \text{TC}(\mathbf{y}, \mathbf{c}, \mathbf{C}) = \sum_{i=1}^N C_{c_i, y_i}^{(i)}$$

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

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

- ❖ 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

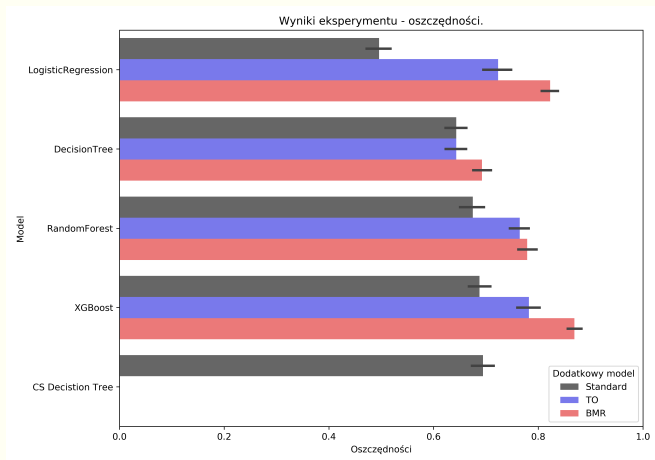
Wyniki

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*).

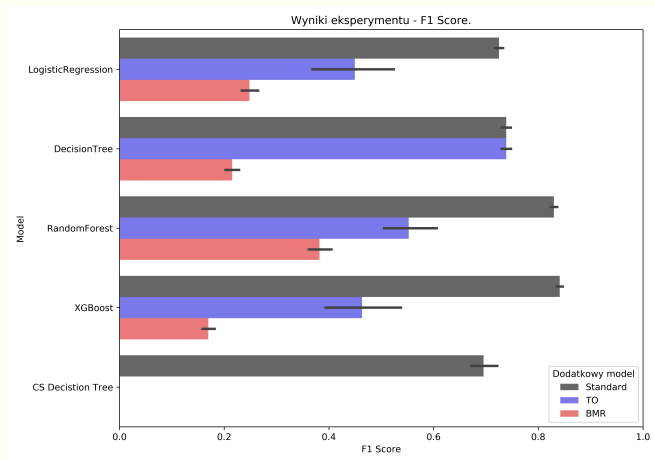
- ❖ 50 powtórzeń symulacji Monte Carlo
- ❖ Podział zbioru danych:
 - ❖ 50% zbiór treningowy
 - ❖ 17% zbiór walidacyjny
 - ❖ 33% zbiór testowy
- ❖ Wykorzystane modele:
 - ❖ Modele standardowe: regresja logistyczna, drzewo decyzyjne, las losowy, XGBoost
 - ❖ Model wrażliwy na koszt: drzewo decyzyjne wrażliwe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

Wyniki dla oszczędności



Źródło: Opracowanie własne

Wyniki dla F1 Score



Źródło: Opracowanie własne

Wnioski

Dziękuję za uwagę!