# Detekcja oszustw z wykorzystaniem metod wrażliwych na koszt

Patryk Wielopolski

29 listopada 2019

# Rozdział 1

# Wstęp

Tutaj będzie wstęp.

# Rozdział 2

# Wprowadzenie teoretyczne

W tej części zostaną wprowadzone wszelkie potrzebne miary skuteczności modeli oraz modele predykcyjne, które zostaną wykorzystane do przeprowadzenia eksperymentu.

## 2.1 Miary skuteczności modeli

### 2.1.1 Macierz pomyłek

W tej sekcji zdefiniujemy macierz pomyłek.

| | | Predykcja | |
|---|---|---|---|
| | | Oszustwo | Normalna |
| Prawda | Oszustwo | TP | FN |
| | Normalna | FP | TN |

Tabela 2.1: Macierz pomyłek

Na podstawie podanej macierzy pomyłek w tabeli 2.1 definiujemy następujące miary skuteczności modeli:

$$\text{Skuteczność} = \frac{TP + TN}{TP + FP + FN + TN}$$

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

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

$$\text{F1 Score} = 2 \cdot \frac{\text{Precyzja} \cdot \text{Czułość}}{\text{Precyzja} + \text{Czułość}}$$

### 2.1.2 Metryki wrażliwe na koszt

## 2.2 Standardowe modele

### 2.2.1 Regresja logistyczna

Formulation of standard Logistic Regression:

$$\hat{p} = P(y = 1|\boldsymbol{x_i}) = h_\theta(\boldsymbol{x_i}) = g\left(\sum_{j=1}^{k} \theta^{(j)} x_i^{(j)}\right)$$

Where loss function is defined:

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} J_i(\theta)$$

Where:

- $g(z) = \dfrac{1}{(1 + e^{-z})}$
- $J_i(\theta) = -y_i log(h_\theta(\boldsymbol{x_i})) - (1 - y_i) log(1 - h_\theta(\boldsymbol{x_i}))$

Standard costs:

$$J_i(\theta) \approx \begin{cases} 0, & \text{if } y_i \approx h_\theta(\boldsymbol{x_i}), \\ \infty, & \text{if } y_i \approx (1 - h_\theta(\boldsymbol{x_i})). \end{cases}$$

Thus

$$C_{TP_i} = C_{TN_i} \approx 0$$

$$C_{FP_i} = C_{FN_i} \approx \infty$$

### 2.2.2 Drzewo decyzyjne

Standard impurity measures:

- Misclassification: $I_m(\pi_1) = 1 - \max(\pi_1, 1 - \pi_1)$
- Entropy: $I_e(\pi_1) = -\pi_1 \log(\pi_1) - (1 - \pi_1) \log(1 - \pi_1)$
- Gini: $I_g(\pi_1) = 2\pi_1(1 - \pi_1)$

### 2.2.3 Las losowy

### 2.2.4 XGBoost

## 2.3 Cost Sensitive Training

### 2.3.1 Regresja logistyczna wrażliwa na koszt

Actual costs:

$$J_i^c(\theta) = \begin{cases} C_{TP_i}, & \text{if } y_i = 1 \text{ and } h_\theta(\boldsymbol{x_i}) \approx 1, \\ C_{TN_i}, & \text{if } y_i = 0 \text{ and } h_\theta(\boldsymbol{x_i}) \approx 0, \\ C_{FP_i}, & \text{if } y_i = 0 \text{ and } h_\theta(\boldsymbol{x_i}) \approx 1, \\ C_{FN_i}, & \text{if } y_i = 1 \text{ and } h_\theta(\boldsymbol{x_i}) \approx 0. \end{cases}$$

Cost sensitive loss function:

$$J^c(\theta) = \frac{1}{N}\sum_{i=1}^{N}\Big(y_i\Big(h_\theta(\boldsymbol{x_i})C_{TP_i} + (1 - h_\theta(\boldsymbol{x_i}))C_{FN_i}\Big) + (1 - y_i)\Big(h_\theta(\boldsymbol{x_i})C_{FP_i} + (1 - h_\theta(\boldsymbol{x_i}))C_{TN_i}\Big)\Big)$$

### 2.3.2 Drzewo decyzyjne wrażliwe na koszt

Cost Sensitive impurity measure:

- $I_c(\mathcal{S}) = min\left\{Cost(f_0(\mathcal{S})), Cost(f_1(\mathcal{S}))\right\}$

Where:

- $\pi_1 = \frac{|\mathcal{S}_1|}{|\mathcal{S}|}$ - percentage of positive class
- $\mathcal{S}$ - set of samples

## 2.4 Cost Dependent Classification

### 2.4.1 Optymalizacja progu

### 2.4.2 Bayesian Minimum Risk

Risk associated with predictions:

$$R(p_f|x) = L(p_f|y_f)P(p_f|x) + L(p_f|y_l)P(y_l|x)$$
$$R(p_l|x) = L(p_l|y_l)P(p_l|x) + L(p_l|y_f)P(y_f|x)$$

Classification threshold:

$$R(p_f|x) \leqslant R(p_l|x)$$

Where:

- $P(p_f|x)$, $P(p_l|x)$ - estimated probability of fraud/legimate transaction
- $L(p_i|y_j)$ and $i, j \in \{l, f\}$ - loss function

Exact formula:

$$P(p_f|x) \geqslant \frac{L(p_f|y_l) - L(p_l|y_l)}{L(p_l|y_f) - L(p_f|y_f) - L(p_l|y_l) + L(p_f|y_l)}$$

After reformulation:

$$p \geqslant \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP} - C_{TN} + C_{FP}}$$

# Rozdział 3

# Eksperyment

Celem eksperymentu jest zbadanie jaki wpływ mają na miarę F1 oraz oszczędności mają poszczególne algorytmy.

Do eksperymentu zostanie wykorzystany zbiór danych Credit Card Fraud Detection zawierający 284,807 transakcji w tym zaledwie 492 oszustw. Tabela składa się z 30 kolumn, w tym 28 z nich są to nienazwane, zanonimizowane zmienne, które były wcześniej poddane transformacji PCA (*ang. Principal Component Analysis*), dodatkowo posiadamy informacje dot. czasu transakcji oraz kwoty.

Rozkład kwoty...

Eksperyment został przeprowadzony w następujący sposób: 50-krotnie dzielimy zbiór danych w proporcjach 50:17:33 na zbiór treningowy, walidacyjny oraz testowy. Następnie uczymy wszystkie modele na zbiorze treningowym. Dla modelu XGBoost wykorzystujemy zbiór walidacyjny do procesu wczesnego zatrzymywania (*ang. Early stopping*), natomiast dla modeli BMR oraz TO korzystamy z tego zbioru jako zbiór treningowy. Następnie dla wszystkich modeli dokonujemy predykcyji na zbiorze testowym i mierzymy skuteczność typowań.

# Rozdział 4

# Rezultaty

# Rozdział 5

# Podsumowanie