Fraud detection using Cost Sensitive methods

Patryk Wielopolski

Wrocław Univeristy of Technology and Science
Koło Naukowe Statystyki Matematycznej „Gauss”

Fraud

Fraud - the crime of getting money by deceiving people.

Examples of frauds:

- Using a stolen credit card
- Overstating the cost of compensation
- Reporting event which never happened
- ...

Fraud detection techniques

Fraud detection techniques:

- Probability models
- Anomaly detection
- Data mining
- Expert rules systems
- Pattern recognition
- Machine learning

Problems

Problems which occurs during modeling:

- Insufficient standard metrics
- Highly imbalanced dataset (class disproportion up to 1:1000)
- Labels distortions

Classification of methodology for classification problems

Standard methodology

Standard methodology for classification problem:

- Standard models:
    - Logistice regression
    - Decision Tree
    - Random Forest
    - XGBoost
- Standard metrics:
    - Accuracy
    - Precision
    - Recall
    - F1 Score

Confusion matrix

		Prediction	
		Fraud	Non-Fraud
True	Fraud	TP	FN
	Non-Fraud	FP	TN

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

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

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

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Cost Sensitive Methodology

Cost Sensitive Methodology:

- Cost dependent classification:
  - Threshold optimization
  - Bayesian Minimum Risk
- Cost sensitive training:
  - Cost Sensitive Logistic Regression
  - Cost Sensitive Decision Tree
- Cost sensitive metrics
  - Total cost
  - Savings

Cost matrix

		Prediction	
		Fraud	Non-Fraud
True	Fraud	C_{TP_i}	C_{FN_i}
	Non-Fraud	C_{FP_i}	C_{TN_i}

$$\mathrm{Cost}(f(\boldsymbol{x}_i^*)) = y_i(c_i C_{TP_i} + (1 - c_i) C_{FN_i}) + (1 - y_i)(c_i C_{FP_i} + (1 - c_i) C_{TN_i})$$

- $\boldsymbol{x}_i^* = [\boldsymbol{x}_i, C_{TP_i}, C_{FP_i}, C_{FN_i}, C_{TN_i}]$ - features vector for i-th observation extended by classification cost
- $C_.$ - classification cost for i-th observation
- $f(\cdot)$ - predictive model
- y_i - true label for i-th observation
- c_i - prediction for i-th observation

Cost Sensitive metrics

Cost sensitive metrics:

$$\text{Total cost}(f(\boldsymbol{S})) = \sum_{i=1}^{N} \text{Cost}(f(\boldsymbol{x}_i^*))$$

$$\text{Savings} = \frac{\text{Cost}_l(\boldsymbol{S}) - \text{Cost}(f(\boldsymbol{S}))}{\text{Cost}_l(\boldsymbol{S})}$$

- $\boldsymbol{S}$ - data set
- $\text{Cost}_l = min\{\text{Cost}(f_0(\boldsymbol{S}), \text{Cost}(f_1(\boldsymbol{S})\}$
- $f_a(\boldsymbol{S}) = \boldsymbol{a}$ where $a \in \{0, 1\}$

Cost Sensitive Modeling

Threshold Optimization

We are looking for threshold th such that

$$\underset{th \in [0,1]}{\arg\min}\ \text{Total cost}(\boldsymbol{C}, \boldsymbol{P})$$

Where

- $\boldsymbol{C} = (c_i)_{i=1}^{n}$ - vector of true labels
- $\boldsymbol{P} = (p_i > th)_{i=1}^{n}$ - vector of binary outcomes

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:

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

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

Bayesian Minimum Risk

Exact formula:

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

After reformulation:

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

Cost Sensitive Logistic Regression

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:

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

Cost Sensitive Logistic Regression

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

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

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

Cost Sensitive Logistic Regression

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:

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

Cost Sensitive Decision Tree

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: $I_g(\pi_1) = 2\pi_1(1 - \pi_1)$

Cost Sensitive impurity measure:

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

Where:

- $\pi_1 = \frac{|\mathcal{S}_\mathbf{1}|}{|\mathcal{S}|}$ - percentage of positive class
- $\mathcal{S}$ - set of samples

Experiment

Experiment description

Data set:

- Credit Card Fraud Detection Dataset
- 284,807 transactions with 492 frauds
- Class imbalance approx. 1:600 (0.172% fraud transactions)

Data split:

- Training: 50%
- Validation: 17%
- Test: 33%

Cost matrix for the experiment

Cost matrix:

		Prediction	
		Fraud	Non-Fraud
True	Fraud	$C_{TP_i} = C_a$	$C_{FN_i} = \text{Amt}_i$
	Non-Fraud	$C_{FP_i} = C_a$	$C_{TN_i} = 0$

- Amt_i - Value of transaction
- C_a - Administrative cost

Decision threshold for Bayesian Minimum Risk:

$$p \geq \frac{C_a}{\text{Amt}_i}$$

Results - Savings

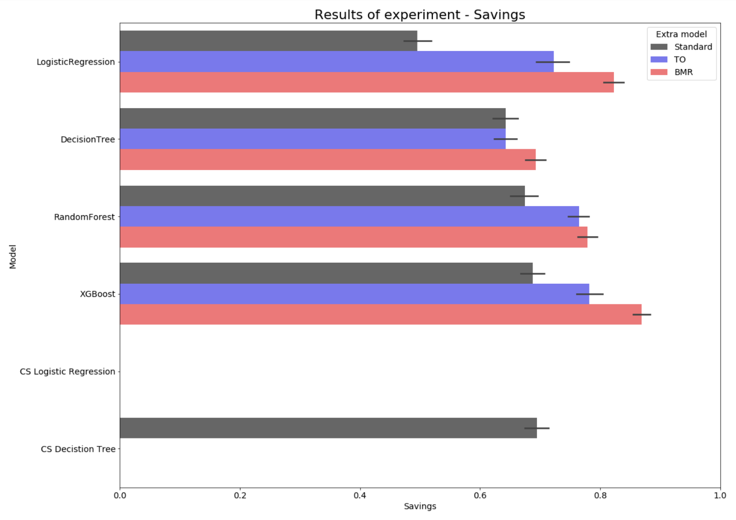

Figure 1: Results of the experiment for $C_a = 1$ and Saving metric.

Results - F1 Score

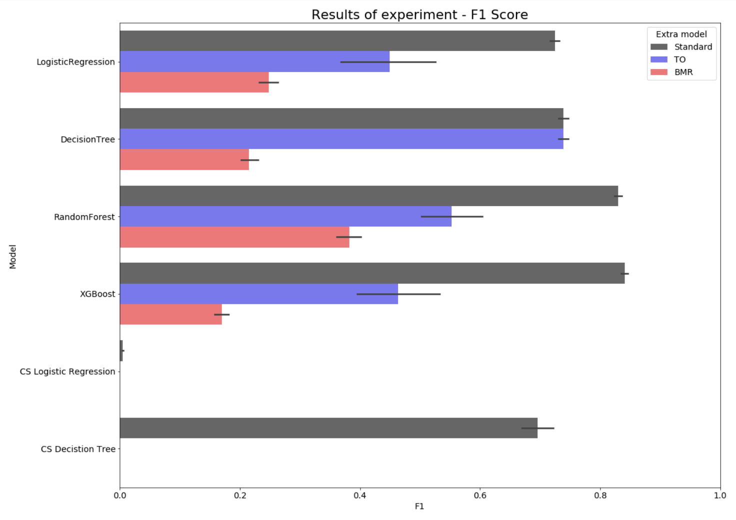

Figure 2: Results of the experiment for $C_a = 1$ and F1 Score metric.

Conclusions

Conclusions:

- Threshold Optimization and Bayesian Minimum Risk improves results for Savings metric,

Conclusions

Conclusions:

- Threshold Optimization and Bayesian Minimum Risk improves results for Savings metric,
- But unfortunately almost always worsens the results for F1 Score metric,

Conclusions

Conclusions:

- Threshold Optimization and Bayesian Minimum Risk improves results for Savings metric,
- But unfortunately almost always worsens the results for F1 Score metric,
- And this leads to trade-off between cost and precision/recall.

Conclusions

Conclusions:

- Threshold Optimization and Bayesian Minimum Risk improves results for Savings metric,
- But unfortunately almost always worsens the results for F1 Score metric,
- And this leads to trade-off between cost and precision/recall.
- Cost Sensitive training may be good choice when Savings and F1 Score are simultaneously important.

Future work

Future work:

- Conducting the experiment in respect to different administrative cost.
- Extension of the used models with Cost Sensitive Ensembles and other boosting models (e.g. LightGMB, CatBoost).
- Extension of the experiment with under/over-sampling methods.
- Defining custom loss function (similarly to Logistic Regression) for boosting algorithms.

References i

A. C. Bahnsen, D. Aouada, and B. Ottersten.
Ensemble of example-dependent cost-sensitive decision trees, 2015.

A. C. Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten.
***Improving Credit Card Fraud Detection with Calibrated Probabilities*, pages 677–685.**

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

Thanks for your attention!

Questions?