Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring

Alejandro Correa Bahnsen, Djamila Aouada and Björn Ottersten

Interdisciplinary Centre for Security, Reliability and Trust - University of Luxembourg

{alejandro.correa, djamila.aouada, bjorn.ottersten}@uni.lu

securityandtrust.lu

Abstract

Several real-world classification problems are example-dependent cost-sensitive in nature, where the costs due to misclassification vary between examples. Credit scoring is a typical example of cost-sensitive classification. However, it is usually treated using methods that do not take into account the real financial costs associated with the lending business. In this paper, we propose a new example-dependent cost matrix for credit scoring. Furthermore, we propose an algorithm that introduces the example-dependent costs into a logistic regression. Using two publicly available datasets, we compare our proposed method against state-of-the-art example-dependent cost-sensitive algorithms. The results highlight the importance of using real financial costs. Moreover, by using the proposed cost-sensitive logistic regression, significant improvements are made in the sense of higher savings.

Cost-Sensitive Evaluation Measure

	Actual Positive $y_i = 1$	Actual Negative $y_i = 0$
Predicted Positive $c_i = 1$	$C_{TP_i} = 0$	$C_{FP_i} = r_i + C^a_{FP}$
Predicted Negative $c_i = 0$	$C_{FN_i} = Cl_i \cdot L_{gd}$	$C_{TN_i} = 0$

Where: r_i is the loss of profit by rejecting customer i, CL_i is the credit limit of customer i, L_{gd} is the loss given default and $C^a{}_{FP}$ is expected income of an alternative borrower

$$Cost(f(S)) = \sum_{i=1}^{N} \Big(y_i(c_i C_{TP_i} + (1-c_i)C_{FN_i}) + (1-y_i)(c_i C_{FP_i} + (1-c_i)C_{TN_i}) \Big)$$

$$Savings(f(S)) = \frac{Cost(f(S)) - Cost_l(S)}{Cost_l(S)}$$

Cost-Sensitive Logistic Regression

Logistic Regression

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

Cost function

$$J_i(\theta) = -y_i \log(h_\theta(X_i)) - (1-y_i)\log(1-h_\theta(X_i))$$

Analysis of the cost function

$$J_i(\theta) \approx \begin{cases} 0 & \text{if} \quad y_i \approx h_\theta(X_i) \\ \inf & \text{if} \quad y_i \approx (1-h_\theta(X_i)) \end{cases}$$

Implicit costs for the different outcomes

$$C_{TP_i} = C_{TN_i} \approx 0 \qquad C_{FP_i} = C_{FN_i} \approx \inf$$

Including the real financial example-dependent costs

Actual costs per example

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

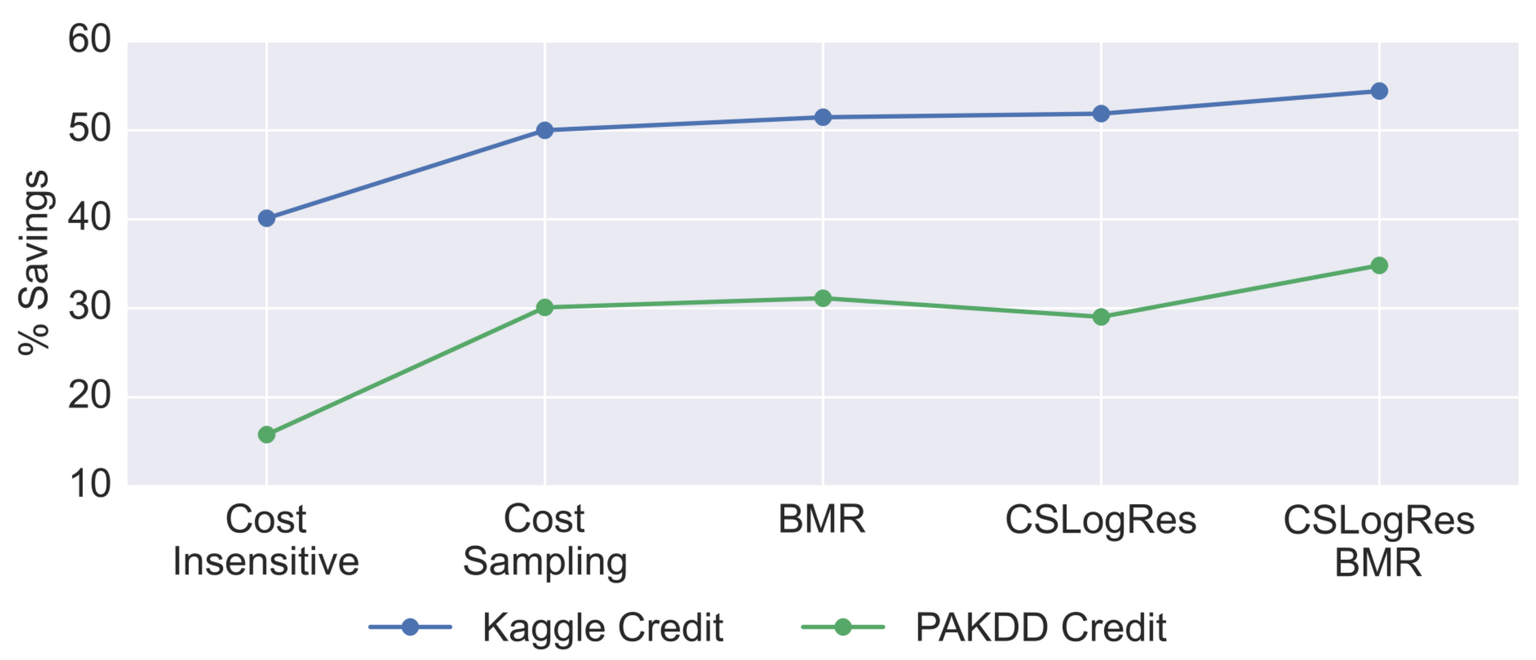
Proposed example-dependent cost-sensitive cost function

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

Experiments

Two credit scoring datasets → Decision Trees / Logistic Regression / Random Forest → Cost-Proportionate Rejection-Sampling / Over-Sampling → Bayes Minimum Risk (BMR) → Cost-Sensitive Logistic Regression

Results



Conclusions

- Selecting models based on traditional statistics does not give the best results in terms of cost
- Models should be evaluated taking into account real financial costs of the application
- Algorithms should be developed to incorporate those financial costs

This project is supported by:

Fonds National de la Recherche Luxembourg

Source code:

https://github.com/albahnsen/CostSensitiveClassification

UNIVERSITÉ DU LUXEMBOURG