International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST

Dependent Cost-Sensitive Credit Cards Fraud Detection Using SMOTE and Bayes Minimum Risk

V Samba Siva[1], Sreya MK[2], Sai Eshmita G[3], Sudheer Kumar M[4], Yugandhar L[5]

[1]Assistant Professor, Department of CSE Faculty, Annamacharya Institute of Technology And Sciences , Tirupati

[2,3,4,5]Students, Department of CSE Faculty , Annamacharya Institute of Technology And Sciences , Tirupati

ARTICLEINFO

Article History:

Accepted: 03 March 2024
Published: 28 March 2024

Publication Issue :
Volume 11, Issue 11
March-April-2024

Page Number :
573-578

ABSTRACT

This paper presents fraud detection problem as one of the most common problems in secure banking research field, due to its importance in reducing the losses of banks and e-transactions companies. Our work will include: applying the common classification algorithms such as logistic regression (LR), random forest (RF), alongside with modern classifiers with state-of-the-art results as XGBoost (XG) and CatBoost (CB), testing the effect of the unbalanced data through com paring their results with and without balancing, then focusing on the savings measure to test the effect of cost-sensitive wrapping of Bayes minimum risk (BMR), we will concentrate on using F1-score, AUC and Savings measures after using the traditional measures duo to their suitability to our problem. The results show that CB has the best savings (0.7158) alone, (0.971) when using SMOTE and (0.9762) with SMOTE and BMR, while XG has the best savings (0.757) when using BMR without SMOTE.

Keywords: Machine learning, Example-dependent cost-sensitive, Random forest (RF), Extreme gradient boosting (XGBoost-XG), CatBoost (CB), Synthetic minority over-sampling technique (SMOTE), Bayes minimum risk (BMR), Fraud detection (FD)

I. INTRODUCTION

From the beginning of the monetary transactions, the fraudsters have tried to gain money in multiple illegal ways, so using protection methods was a necessity. The communications development and moving towards electronic monetary transactions make the fraud more common specially with the ease of exchanging expe riences between the fraudsters and gaining access to the victim companies. The huge losses of banks and other

financial institu tions caused the increase of interest in research to pre vent fraud and decrease its effects. However, methods and techniques could not be revealed to the public, because of the privacy imposed by the supporting companies of these researches, one reason is the high competition in the field, the other is to make sure that fraudsters cannot benefit from the results in improving their methods. For the same reasons there was no standard dataset for research until 2015 when researchers published the fraud detection dataset [1]. Many researchers worked in this field and still, not only to solve a scientific problem but to help real companies and financial institutions to reduce their daily losses. However, some of them used statistics meanwhile oth ers used machine learning approaches, supervised or unsupervised.

1 Recent Works

The high importance of this problem urges the research ers to pay attention to find solutions, that was by using existing methods or developing new ones. In [10] the In [18] they altered the cost function of SVM to produce a cost-sensitive version and they trained it on 21 datasets from KDD98 not including fraud detection dataset, they even compare the results with balancing using SMOTE, they compared the results using AUC (and risk for only datasets with costs included) and their proposed algo rithm had the best results in most of the datasets, mean while in [19] the researchers propose a cost-sensitive ran dom forest based ensemble learning technique and their algorithm outcome two existing cost-sensitive implemen tation of random forest. Although [20] have altered three boosting classifiers to be cost-sensitive it didn't use XGBoost nor CatBoost, how ever the researchers managed to get better results using F-score and Cost as measures. In this paper, we will compare the example-dependent cost-sensitive BMR wrapping of four algorithms (LR, RF, Xgboost and CatBoost) with and without using SMOTE as a rebalancing pre-process step, while using the F1-score, AUC and Savings (the latter is implemented manually) measures. Table 3 shows a comparison between the related and our work.

3 Proposed Work Explanation

3.1 Binary Classification

As we already mentioned, our problem is supervised binary classification, where the dataset includes examples, each is consisted of input and output to train the model, and predict the output of a new example by having the input features. In our work we will use $y \in \{0,1\}$ to refer to the output, where "1" means fraud and "0" means not a fraud.

Table 1 Binary confusion matrix [21]

		True class	
		p	n
Hypothesized class	Y	True positives	False negatives
	N	False negatives	True negatives
	Column totals	P	N

Table 2 Example-dependent cost-sensitive matrix

		True class	
		$y_i = 1$	$y_i = 0$
Hypothesized	$c_i = 1$	C_a	C_a
	$c_i = 0$	Amt_i	0

FIG 1: Binary Confusion Matrix

3.2 Proposed Methodology Architecture

In our work we used the OSEMN process shown in Fig. 2 described in [24] which consists of five steps: Obtain, Scrub, Explore, Model and iNterpret. For the training phase the Obtain was from an CSV file (the dataset) meanwhile, the prediction was from a stream (NiFi simulation). The Scrub has only the scaling (there was no null values in the data and we did not delete outliers for their importance). In the Explore we tested the correlation between the fea tures, and between them and the class. The Model, it has the model parameters setting (with/without SMOTE and with/without BMR) and the cross validation, and in the end the iNterpret, compares one or more model with the F1-score, AUC and Savings measures. Figure 3 shows the details of all the previous steps. In this paper, we will discuss four main algorithms (LR, RF, Xgboost, CatBoost), which some of them is sensi tive to unbalanced data (ex: RF) and the others are not, then we will combine them with data-level sampling algorithm (SMOTE) and finally, we will wrap them with and predict the output of a new example by having the input features. In our work we will use $y \in \{0,1\}$ to refer to the output, where "1" means fraud and "0" means not a fraud.

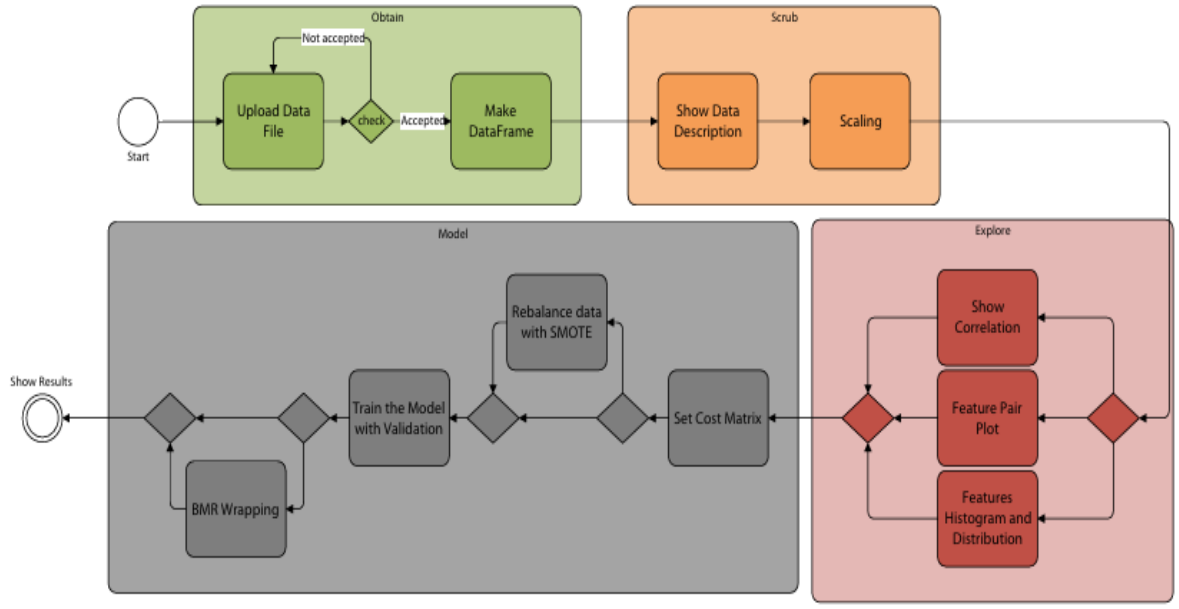

FIG 2: OSEMN in our work.

3.3 Evaluation Metrics

These metrics are commonly used in classification problems and even in FD as in, but they are not preferred in our problem because the data is unbalanced, therefore any classifier, even random classifier, will give a high accuracy if it classifies all the transactions as not fraud. It is more accurate to use other measures as ROC (receiving operating characteristic), AUC (Area Under the ROC Curve) in this type of problems, where ROC curve draws the relation between True Positive percent and False Positive percent and the AUC measure the area under this curve, where the higher AUC the better.
most suitable for evaluation in our problem. hence, we can use the proposed cost matrix in which is example-dependent. And from this matrix we can calculate the cost and savings respectively as follows:

Cost = $\sum N\ i=1\ yi\ (\ 1 - ci\ )\ Amti + ci\ Ca$ …
Savings = 1 – Cost/ Costl ,…
where Cost (3) l = $\sum N\ i=1\ yi\ Amti$ …
N: the number of examples.
Costl : the cost of not using any algorithm and predict all the examples as not fraud.

FIG 3: OSEMN process

4 Conclusions

In this paper we studied the fraud detection problem in credit cards, presenting the methods to reduce the unbal ancing of the data using resampling SMOTE as a preproc ess. We compare some common classifiers with and with out cost-sensitive wrapping by F1-score, AUC and Savings measures. As mentioned before the main challenges in our problem is the unbalanced data and the concept drift, in this paper we were concerned with the frst challenge, meanwhile the dataset cannot be used to study the second one, due to the independence of the transactions. As we already mentioned the dataset was altered for privacy reasons by deleting the id of the credit card so we cannot connect two or more transaction belongs to the same credit card. In addition, we scaled the Time and Amount features so they will have the same efect as the others, but we used the Amount to weight the risk in BMR wrapping and again in the Savings measure. Finally, we found that XG has given good Sav ings when wrapped with BMR, but CB and RF has outper formed when using SMOTE. As future work, we can consider testing XG and CB for example-dependent cost-sensitivity by modifying their loss function as future work, which consider during train ing cost-sensitive implementation.

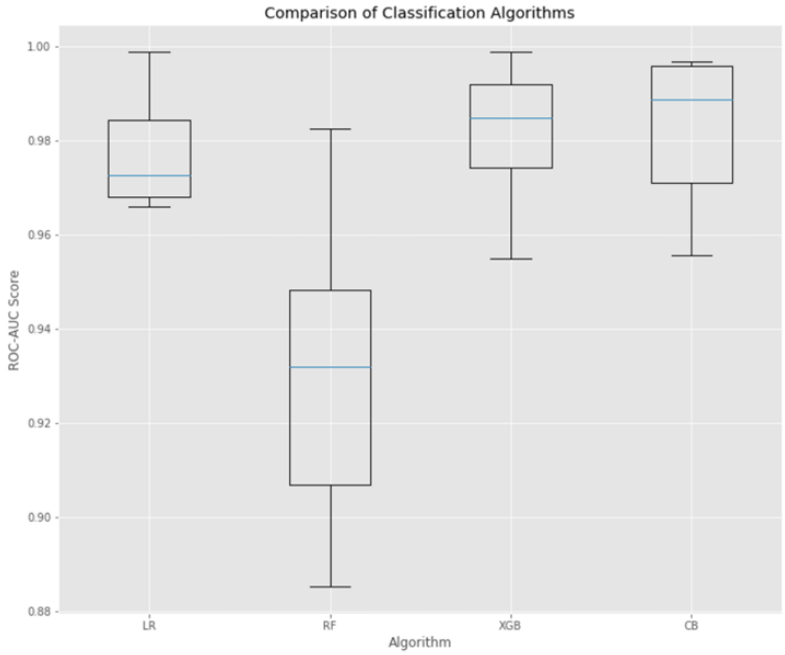

FIG 4: AUC comparison of the classifiers.

References

1. Dal Pozzolo A, Bontempi G (2015) Adaptive machine learning for credit card fraud detection. Unpublished doctoral dissertation, Université libre de Bruxelles, Faculté des Sciences—Informa tique, Bruxelles.
2. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
3. Devi D, Biswas SK, Purkayastha B (2019) A Cost-sensitive weighted random forest technique for credit card fraud detec tion. In: 2019 10th international conference on computing, com munication and networking technologies (ICCCNT). IEEE, pp 1–6
4. Sharifnia E, Boostani R (2020) Instance-based cost-sensitive boosting. Int J Pattern Recognit Artif Intell 34(03):2050002
5. Fawcett T (2006) An introduction to ROC analysis. Pattern Rec ognit Lett 27(8):861–874
6. Dhankhad S, Mohammed E, Far B (2018) Supervised machine learning algorithms for credit card fraudulent transaction detec tion: a comparative study. In: 2018 IEEE international conference on information reuse and integration (IRI). IEEE, pp 122–125
7. Bahnsen AC, Aouada D, Stojanovic A, Ottersten B (2016) Feature engineering strategies for credit card fraud detection. Expert Syst Appl 51:134–142
8. Park Y, Luo L, Parhi KK, Netoff T (2011) Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. Epilepsia 52(10):1761–1770

Authors

Mr. V Samba Siva., M.Tech(Ph.d), is currently working as Assistant Professor in the department of CSE, Annamacharya Institute of Technology and Sciences, Tirupati.

Ms. MK Sreya is currently pursuing B.Tech in the stream of Computer Science and Engineering from Annamacharya Institute of Technology and Sciences, Tirupati.

Ms. G Sai Eshmita is currently pursuing B.Tech in the stream of Computer Science and Engineering from Annamacharya Institute of Technology and Sciences, Tirupati.

Mr. M Sudheer Kumar is currently pursuing B.Tech in the stream of Computer Science and Engineering from Annamacharya Institute of Technology and Sciences, Tirupati.

Mr. L Yugandhar is currently pursuing B.Tech in the stream of Computer Science and Engineering from Annamacharya Institute of Technology and Sciences, Tirupati.