Credit Card Fraud Detection Based on Multiple Machine Learning Models

Muyuan Chen
Goergen Institute for Data Science, University of Rochester, Rochester, New York
Americamchen68@u.rochester.edu

ABSTRACT

Credit cards have long been one of the most popular methods of making payments. But this kind of payment still carry risks —— credit card fraud. Machine learning has played an essential role in detecting credit card fraud. This study uses machine learning methods to provide most accurate prediction of fraudulent transactions. These algorithms include Random Forest, Logistic Regression, SVM, and Extreme Gradient Boosting (XGBoost). Dataset used in this study is from Kaggle and is highly skewed. In order to find out if imbalanced dataset will affect predictions of models, three oversampling methods, Random oversampling, SMOTE, and ADASYN, are used to resample dataset for comparative analysis. From the analyzed results, oversampling methods tend to cause overfitting of models. XGBoost is the only technique that are not affected by oversampling methods. Since overfitting is a problem that should be avoided in classification problem, original data is selected for further evaluation of models. Performance of four algorithms are evaluated based on F-measure, Cohen's Kappa, AUC score, and accuracy. The results show that Random Forest and XGBoost outperform SVM and Logistic Regression. As a result, a fusion model is constructed based on these two classifiers. Log loss and brier score are used to further evaluate their predictions. The results show that Random Forest and XGBoost provide best predictions and close to each other. Fusion model has a little improvement compared to these two classifiers.

CCS CONCEPTS

• **Networks, Information systems**;

KEYWORDS

Credit card fraud detection, Logistic Regression, XGBoost, Oversampling methods

ACM Reference Format:
Muyuan Chen. 2022. Credit Card Fraud Detection Based on Multiple Machine Learning Models. In *2022 6th International Conference on Electronic Information Technology and Computer Engineering (EITCE 2022), October 21–23, 2022, Xiamen, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3573428.3573745

EITCE 2022, October 21–23, 2022, Xiamen, China

ACM ISBN 978-1-4503-9714-8/22/10…$15.00
https://doi.org/10.1145/3573428.3573745

1 INTRODUCTION

The outbreak of COVID-19 in early 2020 has temporarily changed people's lives in different perspectives. The way of payment is one of them. Due to the pandemic, payment preferences have changed to more contactless payments, which caused a sharp rise in the volume of credit card transactions [1].

As people rely more on payments using credit card, fraudulent transaction has become a significant problem. Recent information from the Nilson Report indicates that merchant losses from card-not-present fraud increased six times in 2020 compared to only one year earlier [2]. Not only do consumers need to be concerned about fake cards. In fact, an increasing number of fraudsters choose to completely disregard physical cards and instead use sophisticated scams to divulge detailed cardholders' information. With these information, they can take control of already-existing accounts or establish new ones on behalf of fictitious individuals. This is what people call card-not-present, or eCommerce fraud. Nowadays, fraudulent transactions have a big impact on both businesses and customers. The ability to identify fraud has improved, but fraudsters are always searching for new ways to avoid being caught.

Capturing unusual transaction patterns is a key step in fraud detection, and accurate detection of credit card fraud requires efficient and effective classification techniques [3]. Basically, these methods can be categorized into two groups, namely unsupervised and supervised learning models. Moreover, imbalanced data is a common problem in fraud detection. Datasets with an uneven distribution of observations in the target class are referred to as imbalanced data. In reality, one could anticipate that there are many more non-fraudulent transactions than fraudulent ones. Thus, dealing with an imbalanced dataset is a challenge in the evaluation of different fraud detection models.

From an earlier study, there are two methods for detecting unsupervised credit card fraud. These two techniques are called peer point analysis and break point analysis [4]. To be more specific, peer group analysis (PGA) analyzes the transaction behavior of an individual account (i.e. target) and finds the accounts that are most similar to the target account. All these accounts form the peer group of any individual account. Then PGA compares the target's behaviors with other accounts to identify whether the transaction is fraudulent or not [4]. As a result, PGA is capable of detecting a new type of fraud. However, this method is not realistic since it is time-consuming and credit card fraud detection should be finished immediately when a transaction happens. Another approach for identifying unsupervised fraud is Break Point Analysis (BPA). Break Point is the moment when abnormal conduct is discovered [4]. By comparing sequences of an account's transactions, BPA can detect unexpected modification of account behavior over a period of time. Transactions that suddenly rise in frequency or volume might be

viewed as fraudulent activity. The advantage of this method is that it does not need balanced data for assessment because compared transactions are not from different accounts. However, the tests are less potent in the Break Point Analysis than they are in peer group analysis since the profile is not formed using transactions from similar accounts.

Both of these two unsupervised fraud detection models have drawbacks. Many supervised models have also been used to identify fraudulent transactions. In this study, support vector machines and random forests, two data mining techniques, are evaluated [5]. This research is designed to overcome the limitations of both algorithms and provide a more exact and accurate value by combining their strengths to identify fraud in the available dataset [5]. Both two algorithms have been proven to be effective in fraud detection. However, their evaluation is based on accuracy for predicting fraudulent cases. It is more reasonable to draw on a more comprehensive conclusion of different models' performance with more indicators.

This paper examines naïve Bayes, logistic regression, and k-nearest neighbors with different data distributions, which are achieved by hybrid sampling [6]. When utilizing the resampled dataset, the naïve Bayes model appears to have superior performance. For the logistic regression model, the original dataset yields better results than do the two re-sampled datasets. The purpose of this research is to compare the performance of three supervised models and shows how hybrid sampling affects the effectiveness of binary classification of unbalanced data.

Based on past research, this article aims to find the model with the best fraud detection ability. Models used in this research include random forest, logistic regression, SVM, and Extreme Gradient Boosting (XGBoost). XGBoost has been proved to be an effective model for prediction problems as random forest [7]. In order to show a more comprehensive analysis of these models, this study not only focuses on basic indicators like accuracy, recall, precision, F-measure scores but also focuses on the model reliability and prediction probability of each model. Prediction probability is achieved and evaluated using Calibration curve. Furthermore, this research investigates the performance of different combinations of four models by using fusion models.

2 METHOD

2.1 Dataset

Dataset used in this project is from Kaggle [8]. Credit card transactions made by European cardholders in September 2013 over the period of two days are included in the dataset. The dataset is very skewed, with 492 frauds within 284,807 transactions. Since displaying customers' transaction details will lead to the problem of confidentiality, most of the dataset's attributes are generated by Principle Component Analysis (PCA). The dataset contains 31 columns, including time, 28 features generated by PCA, amount, and target variable.

2.2 Preprocessing

Preprocessing procedure focuses mainly on resampling the imbalanced dataset. Standard learning methods that emphasize overall accuracy tend to identify all observations as majority class instances when applied to unbalanced data. However, in fraud detection problem, the minority class of fraud samples is more important. Numerous studies have demonstrated that, for numerous classifiers, a balanced dataset outperforms an unbalanced dataset in terms of overall classification performance [9]. As a result, several methods that deal with imbalanced problems are used: Random Oversampling, Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN).

Random oversampling reduces class imbalance by including a set drawn from the class of minorities. For a collection of minority instances chosen at random, random oversampling method adds replicas of the chosen samples to the dataset, augmenting the initial set [10]. By creating synthetic minority cases close to the actual ones, SMOTE oversamples the minority class [10]. New minority examples are produced by interpolating between instances of the same class. For each original minority example, the same number of synthetic data samples are generated by SMOTE. But this generation does not take surrounding instances into account, increasing the likelihood of overlapping between classes. This issue is solved by the ADASYN technique, which employs a methodical approach to adaptively generate various quantities of synthetic data in accordance with their distributions [10].

2.3 Machine Learning Algorithm

Logistic regression classifier. Logistic regression is based on the concept of probability. The sigmoid function in logistic regression is used to return the probability of a label. This mathematical formula is able to convert predicted values to probabilities. Any real value can be converted into another value within a range of 0 and 1. Below are the sigmoid function (σ) and input (s):

$$\sigma(s) = \frac{1}{1+e^{-s}} \tag{1}$$

$$s = w_0x_0 + w_1x_1 + \ldots + w_nx_n \tag{2}$$

The vector x, which represents the input data, and the best coefficients w are multiplied together. The sum of all these productions is a single value that will be converted by sigmoid function into the final target class. The discrepancy between the anticipated value and the actual value is calculated by a cost function. Another name for the cost function's value is loss or error. Cross entropy, commonly known as log loss, is the cost function used in logistic regression. Optimization methods like gradient descent are investigated to minimize the cost, which can estimate the parameters of the model.

SVM. The SVM approach considers each data point as a point in an n-dimensional space. n is the number of features in the input dataset, and each feature represents a specific coordinate. Problem of fraud detection falls into just two classes. A hyperplane is the optimal choice boundary. The data can be categorized by a variety of hyperplanes. An optimal hyperplane can be selected by computing margin between different classes. The optimal hyperplane is the one that maximizes the margin of the training data.

Random Forest. The algorithm generates models by training multiple decision trees, and then classifies them based on prediction of individual trees. The idea behind decision trees is that yes/no questions are created using dataset attributes, and datasets are continuously split into classes until all of the data points fall into each one. Each tree is constructed by measuring loss function, which

also assesses each split by calculating the percentage of data points belonging to each class before and after the split. Loss functions that are commonly used in decision trees are Gini Impurity and Entropy.

A random forest is made up of a huge number of independent decision trees. Each tree produces an outcome (class prediction) and the most frequent outcome will be selected to be the model's final prediction.

Extreme Gradient Boosting. XGBoost is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. Similar to random forest, GBDT is a decision tree ensemble learning algorithm. Both GBDT and random forest create a model made up of several decision trees. The way the trees are constructed and joined makes a difference. A weighted average of each tree estimation makes up the GBDT's final result. A group of decision trees is trained repeatedly, with each iteration using the previous model's error residuals to fit the new model. XGBoost pushes the limits of computing power for boosted tree algorithms in order to improve the model performance and computational speed.

2.4 Calibration curve and fusion models with prediction probability

Many machine learning models can predict class membership with a probability or probability-like scores. This number provides a way for the model to express their certainty about what class the input data belongs to. By using calibration these scores can be transferred into probabilities and be used more effectively in decision making.

Sometimes, individual models are not able to give best prediction results. Fusion is a way to use the combination of different models to solve classification and regression problems. For classification problems, Fusions perform an average of the per-class probabilities across all the component models and predict the class with the highest probability.

3 EVALUATION AND RESULT ANALYSIS

3.1 Oversampling Methods

Four models are created in this study: Logistic Regression, SVM, Random Forest, and XGBoost. To analyze these models, 30% of the dataset is utilized for validation and testing, while the remaining 70% is used to train these models. For evaluation, F-measure, Cohen's Kappa, Precision, Recall, AUC, and Accuracy are utilized. There is one original dataset and three additional datasets using different oversampling methods. The performance of four models using these four datasets is presented in Figure 1, Figure 2, Figure 3, and Figure 4. From these figures, there is an improvement of model's overall performance on training set from the un-sampled dataset to oversampled datasets. But, on the contrary, overall performance is worse on test sets. XGBoost is the only model that is not affected by oversampling methods.

In conclusion, oversampling methods on a highly imbalanced dataset do improve Logistic Regression, SVM, and Random Forest classifiers' ability of fraud detection. However, it also leads to the overfitting problem of these three models, which is the reason that performance of the models on test set is worse from using un-sampled datasets to oversampled datasets.

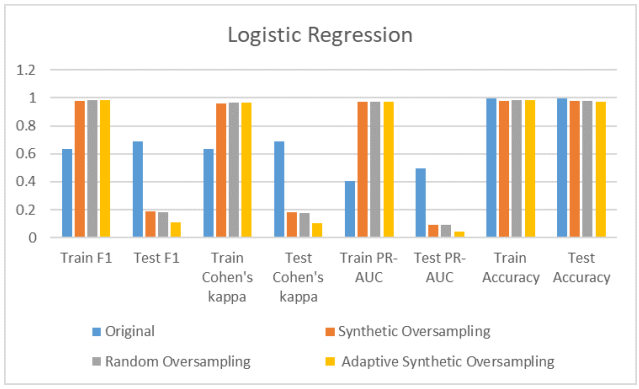

Figure 1: Results of Logistic Regression with four datasets

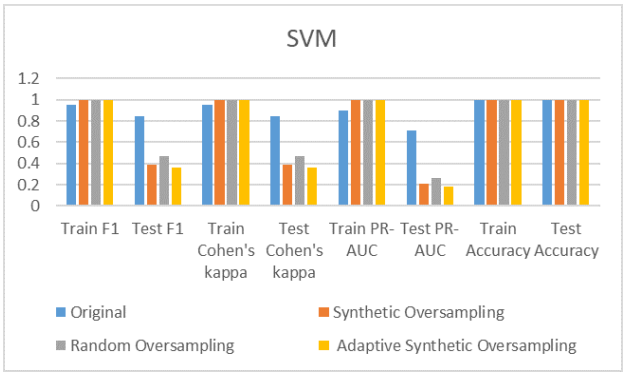

Figure 2: Results of SVM with four datasets

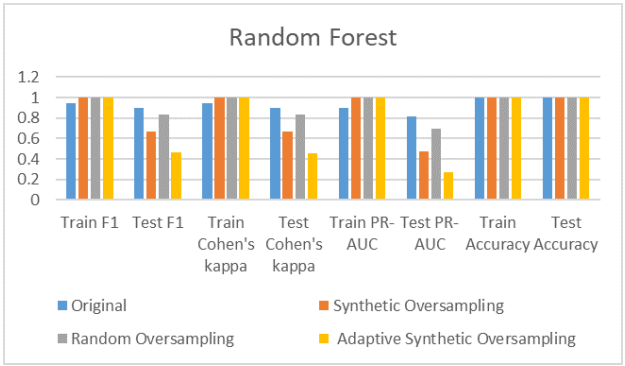

Figure 3: Results of Random Forest with four datasets

3.2 Comparative Analysis

Since oversampled datasets tend to make models to be overfitting, the evaluation of performance is based on un-sampled dataset. Table 1 presents the performance of four models on test set. All four models yielded high accuracies. XGBoost and Random Forest

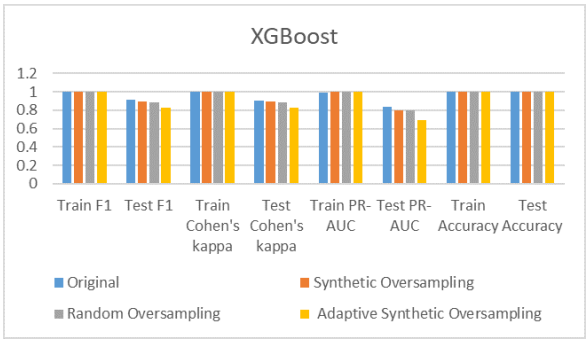

Figure 4: Results of XGBoost with four datasets

showed superior performance. These two tree-structured models outperformed SVM in all of the evaluation standards. Among the four classifiers tested, Logistic Regression performed the worst.

Since Random Forest and XGBoost showed good performance, a fusion model was constructed by using the average of their probabilistic predictions. Predicted probabilities of four models are achieved by using methods of calibration. Table 2 shows the comparative results of five models based on brier score and log loss. Random Forest, XGBoost, and Fusion model showed better performance than Logistic Regression and SVM. There is a slight improvement from XGBoost to Fusion Model. Brier score of Fusion Model is lower than Random Forest, but Log loss of Fusion Model is higher than Random Forest.

From comparative results, high accuracy is obviously not enough to define a good model. According to all evaluation measures, the two top classifiers are Random Forest and XGBoost. They are very close on F-measure, Cohen's Kappa, Precision, Recall, AUC, and Accuracy. Furthermore, brier score and log loss between fusion model and two tree-structured models are close. As a result, it is reasonable to conclude that predictions of Random Forest and XGBoost are stable and close to each other. Fusion model even has a little improvement compared to XGBoost and Random Forest.

4 CONCLUSION

In this paper, different machine learning algorithms are evaluated on resolving the two-class classification problem of credit card fraud detection. This project evaluates performance of Logistic Regression, SVM, Random Forest, and XGBoost using original highly imbalanced dataset based on different evaluation standards. It is established that the two classifiers that produce the best predictions are Random Forest and XGBoost. Extensive experiments are conducted to investigate the effect of oversampling methods to the model prediction. By comparative analysis, oversampling methods tend to cause Logistic Regression, SVM, and Random Forest to overfit, which is a serious problem that should be avoided in classification problems. Such results determine the selection of original dataset that are used for evaluation of machine learning models. Furthermore, a fusion model is constructed based on two best models, Random Forest and XGBoost, to investigate their probabilistic predictions. The results show that predictions provided by Random Forest and XGBoost are close.

ACKNOWLEDGMENTS

Further research should focus on other machine learning algorithms such as unsupervised models and neural networks. Also, more sampling methods can be investigated on their effects on the detection of credit card fraud samples.

REFERENCES

[1] Jonker, N., *et al.* 2022. Pandemic payment patterns. Journal of banking & finance, 143, 106593. https://doi.org/10.1016/j.jbankfin.2022.106593
[2] Nilsonreport, 2022, https://nilsonreport.com/.
[3] Sudhamathy, G. 2016. Credit risk analysis and prediction modelling of bank loans using R. Int. J. Eng. Technol 8.5, 1954-1966.
[4] Weston, D J *et al.* 2008. Plastic card fraud detection using peer group analysis. Advances in Data Analysis and Classification,2, Springer Berlin, 45-62.

Table 1: Performance evaluation of four classifiers

	F1	F2	Cohen's Kappa	AUC	Accuracy
XGBoost	0.9191	0.8621	0.9089	0.8337	0.9996
Random Forest	0.9478	0.8441	0.8963	0.8129	0.9996
SVM	0.8409	0.7974	0.8406	0.7136	0.9993
Logistic Regression	0.6914	0.6222	0.6908	0.4959	0.9988

Table 2: Probabilistic prediction evaluation of five models

	Logistic Regression	SVM	RandomForest	XGBoost	FusionModel
Brier Score	0.00091	0.00049	0.00017	0.00021	0.00015
Log Loss	0.0046	0.0027	0.00087	0.0015	0.0011

[5] Saddam Hussain, S, K., *et al.* 2021. Fraud Detection in Credit Card Transactions Using SVM and Random Forest Algorithms 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 1013-1017
[6] Awoyemi, J, O., 2017, Credit card fraud detection using machine learning techniques: A comparative analysis, 2017 International Conference on Computing Networking and Informatics (ICCNI), pp. 1-9.
[7] Sheridan, R. P., 2016, Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships. Journal of chemical information and modeling, 56(12), 2353–2360.
[8] Kaggle, 2018, https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
[9] Weiss G, W., 2001, The Effect of Class Distribution on Classifier Learning: An Empirical Study, Technical Report MLTR-43, Dept. of Computer Science, Rutgers Univ.
[10] Haibo, H., 2009, Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.