Improved credit card fraud detection method based on XGBoost algorithm

Fanrui Zhang

College of Statistics and Data Science, Nankai University, Tianjin, 300072, China

2014078@mail.nankai.edu.cn

Abstract. With the development of the Internet and technology, credit cards are more widely used and transaction data are larger. The data set of credit card fraud is a typical imbalanced data problem. The model should ensure that fraud is detected and customer service quality is guaranteed. Improving both the precision and the recall rate is the focus of current research. However, when the precision is constrained by the current level of machine learning, it is a good choice to use different models to evaluate and obtain more data for manual use. Few papers use the results of both models for comprehensive judgment. In this paper, two sampling methods (undersampling-NearMiss, oversampling-SMOTE) and three algorithms (Logistic, Neural Network, XGBoost) are used to analyze data based on the transaction records in Europe within two days in September 2013. The above six cases were compared to explore appropriate detection methods. The study found that Nearmiss-XGBoost had the best recall rate, SMOTE -XGBoost had the best comprehensive result and precision. Compounding the results of the two models can improve the precision and recall.

Keywords: Xgboost algorithm; Credit card fraud; Unbalanced data; Smote and Near Miss.

1. Introduction

According to the Nielsen Report, an authoritative research institute on the global payment industry, payment card fraud caused $28.65 billion in business losses worldwide in 2019. The 2019 American Express Digital Payments Survey found that 42 percent of consumers have experienced fraud when trying to use credit cards or other payment information,77 percent of U.S. merchants have experienced fraud in the course of their business, and 59 percent of consumers are concerned about the disclosure of payment account or credit card information for online purchases [1,2]. In recent years, the outbreak of the coronavirus pandemic has fueled explosive growth in credit card fraud. The problem of credit card fraud is widely concerned by consumers, merchants and card issuers [3,4].

There are many researches related to credit card fraud detection model. In order to get good data, each model is subjected to complex sampling methods (undersampling and oversampling, repeated sampling, etc.) and iterative processes [5]. The literatures aim to inform the reader of the construction principle of the credit card fraud detection model and the necessity of each basic step through the construction of the most basic model [6,7]. Most of the current literature uses only one path to build the model (one sampling method + one algorithm) [8,9].

In this paper, two sampling methods (NearMiss and SMOTE) and three algorithms (the more classical logistic regression algorithm are selected, the widely used neural network, and the more cutting-edge XGBoost algorithm) to make a horizontal comparison of the above six combinations. It directly reflects the advantages and disadvantages of different sampling methods and algorithms. Research on Credit Card Application Anti-Fraud Scoring Model of Small and Medium-sized Banks in China -- Based on AHP Method offers a detailed look at the reality of credit card fraud. then found that in real life credit card fraud is judged according to the hierarchy. At the end of the article, a method is put forward to classify the detected results by using two models, and discusses the possibility of improving the effect of the model.

2. Research Methodologies

2.1 Background

Today's credit cards are mainly faced with huge data, fraud accounts for a relatively low, very time-sensitive problems. A good model needs to retrieve a large amount of information in a very short time to screen out fraudulent behaviors [10]. At the same time, facing the vast consumer group, it is necessary to reduce the normal behavior misjudgment, improve the quality of credit card service.

2.2 Data Standardizition and correlation Analysis

The threshold of different independent variables is different. When all variables have the same impact on the dependent variable, the independent variables need to be standardized.

Standardization means changing the threshold interval to [-1, 1]. A common method is the RobustScaler formula

$$x' = \frac{x - x_{min}}{Q_3(x) - Q_1(x)}, Q_i \text{ is the i quantile of x } i = 1,3 \quad (1)$$

Thermal map (correlation coefficient map) is a commonly used method for correlation analysis.

Each value in the figure represents the correlation coefficient

$$\rho_{(xi,xj)} = \frac{Cov(x_i, x_j)}{\sqrt{D(x_i) * D(x_j)}} \quad (2)$$

Cov(xi,xj) is the covariance of xi and xj, D(xi) is the variance of xi.

2.3 2.3 Imbalanced data process

For imbalanced data, it is necessary to ensure the balance between fraudulent and non-fraudulent data in the sampling process.

There are two sampling methods: undersampling and oversampling. Undersampling is the elimination of data from most classes. Oversampling is the generation of new data in a few classes. This paper used NearMiss (undersampling)and SMOTE (oversampling).

2.3.1 NearMiss

(1) find the distance between all instances of the majority class and the instances of the minority class

(2) Select N instances of the majority class with the smallest distance from the minority class

(3) if there are k instances in the minority class, the nearest method will result in K *n instances of the majority class

2.3.2 SMOTE

SMOTE, short for Synthetic Minority Oversampling Technique:

(1) set the set of minority classes A, and for each, obtain the k-nearest neighbors of X by calculating the Euclidean distance between X and every other sample in the set A.

(2) Set the sampling rate N according to the unbalanced proportion. For each, select a random example from its k-nearest neighbors (i.e. X1, x2, ... Xn) and build the collection.

(3) for each example (k=1, 2, 3... N), generate a new sample using the following formula: where rand (0, 1) represents a random number between 0 and 1.

2.4 XGboost

Known training set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, Loss function $l = (y_i, \hat{y_i})$, Regularization item $\Omega(f_k)$.

The overall objective function can be written as

$$L(\phi) = \sum_i l(y_i, \hat{y_i}) + \sum_k \Omega(f_k) \quad (3)$$

$L(\phi)$ is a statement on linear space; i is the ith sample, k is the tree, $\hat{y}_i$is the predicted value of the ith sample。

Compared with GBDT, XGBoost (Extreme Gradient Boosting) makes optimization in the following three aspects

(1) Second order Taylor expansion, removing the constant term (GBDT is a first order Taylor expansion)

$$l(y_i, \hat{y}_i) \approx l(y_i, \hat{y}_i^{(t-1)}) + \acute{l}(y_i, \hat{y}_i^{(t-1)})(x - \hat{y}_i^{(t-1)}) + \frac{1}{2}\ddot{l}(y_i, \hat{y}_i^{(t-1)})(x - \hat{y}_i^{(t-1)})^2 \quad (4)$$

$x = \hat{y}_i^{(t-1)} + f_t(x_i)$, first derivative$g_i = \acute{l}(y_i, \hat{y}_i^{(t-1)})$, second derivative$h_i = \ddot{l}(y_i, \hat{y}_i^{(t-1)})$.

$$l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) \approx l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \quad (5)$$

$$L^{(t)} = \sum_i l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_k \Omega(f_k) \quad (6)$$

Since $l(y_i, \hat{y}_i^{(t-1)})$ is a constant term, delete it.

(2) The regularization term is expanded and the constant term is removed

$$\sum_k \Omega(f_k) = \sum_{k=1}^{t} \Omega(f_k) = \Omega(f_t) + c \quad (7)$$

Delete c.

(3) Combine the coefficients of the first term and the second term

Define:$G_j = \sum_{i \in I_j} g_i$, $H_j = \sum_{i \in I_j} h_i$

$$L^{(t)} = \sum_i [G_i w_j + \frac{1}{2}(H_i + \lambda) w_j^2] + \gamma T \quad (8)$$

For the solution of this function, according to the properties of quadratic functions, when$w_j = -\frac{G_j}{H_j + \lambda}$ has the min value， $\min f(w_j) = -\frac{G_j^2}{2(H_j + \lambda)}$.

3. Experiments

In order to test the effectiveness of the proposed method, in this section, simulation experiments with real data are conduct.

3.1 Data Source

The test data came from two days of European cardholders' transactions in September 2013. There were 284,807 records, including 492 fraudulent transactions. The information of the transaction is desensitized and consists of 28 dimensional vectors. Time column indicates the transaction Time. Class column indicates fraud or not: 1 indicates fraud and 0 indicates normal behavior. Amount column represents the Amount of the transaction.

3.2 The Experiment Design

3.2.1data processing

Since columns V1-V28 are desensitized, their values are standardized. Since the contents of V1-V28 are encrypted, their relative importance with amount and time cannot be verified, so the columns of time and amount are standardized.

After standardized processing, the data sets were divided into training set and test set by 8:2 ratio. The purpose of data segmentation before sampling is to test the model with the original data, which has a better test effect.

3.2.2 Analysis of correlation

Random undersampling is performed on the training set to balance the data. Generate correlation coefficient maps

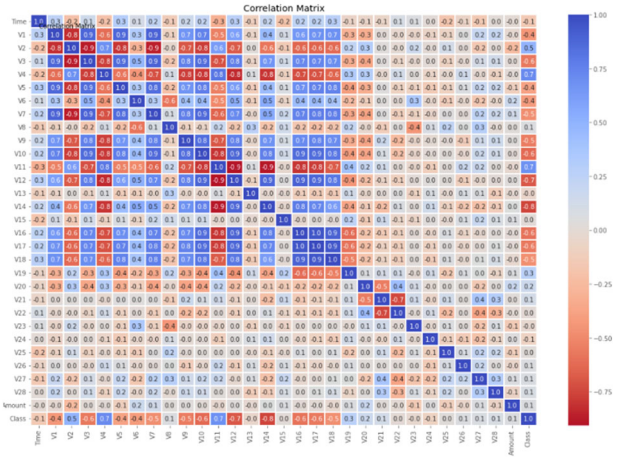


Figure 1. Coefficient maps

V10, V12, V14, V16, V17 are negatively correlated with the dependent variable, while V4, V11 are positively correlated with the dependent variable. Remove extreme outliers. The interquartile interval method is selected to eliminate outliers and visualize them with boxplot. Select a range of 1.5 times the quartile to determine the threshold.

Use t-SNE to reduce the dimension and classifier, detecting the processing of the data.

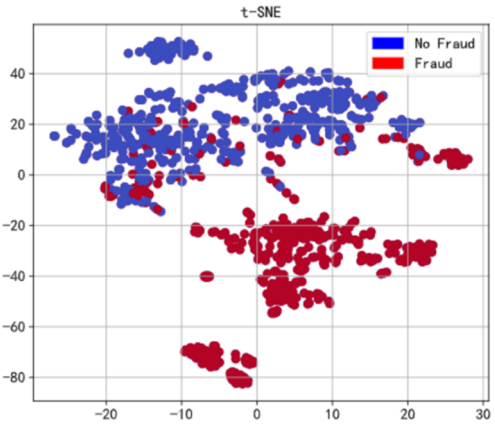
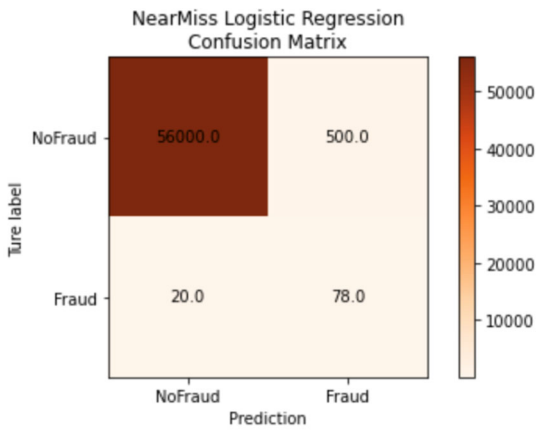


Figure 2 Data visualization

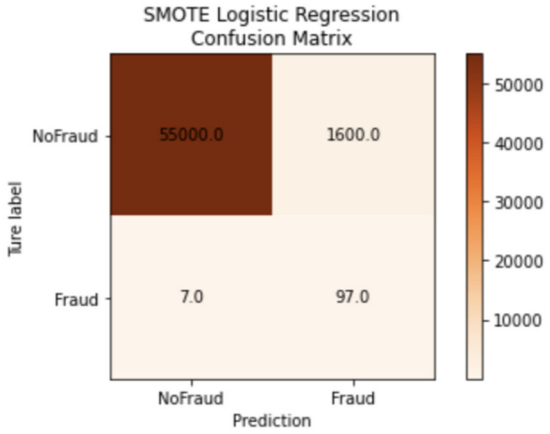
3.2.3 Sampling and methods

Data sampling according to the NearMiss and SMOTE method, respectively. And three classifiers are used in this paper: Logistic Regression, Neural Networks and XGBoost. The advantages and disadvantages of the three algorithms are compared horizontally.

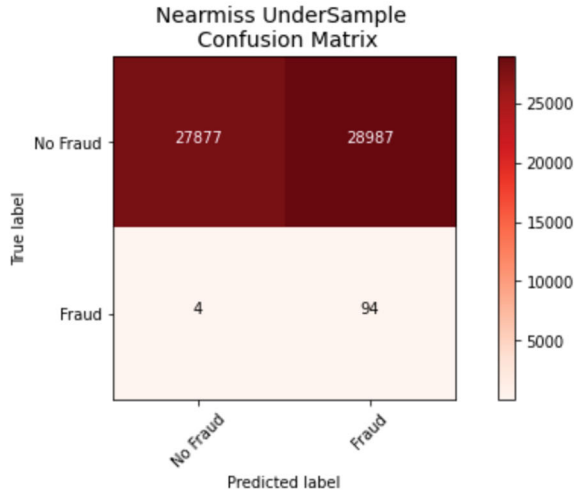
The following is the confusion matrix of the three algorithms



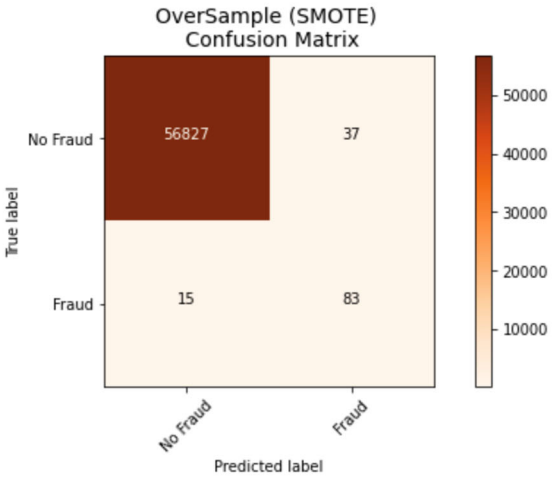
NearMiss-Logistic Regression



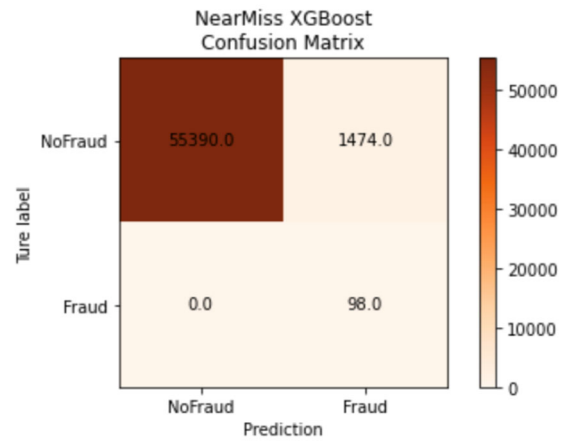
SMOTE-Logistic Regression



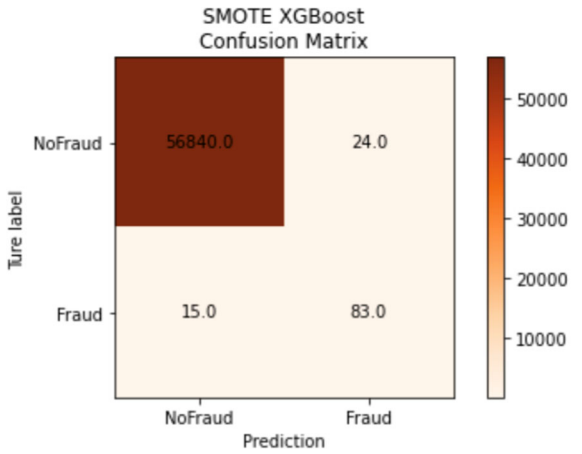
NearMiss-Neural Network



SMOTE-Neural Network



NearMiss-XGBoost



SMOTE-XGBoost

Figure 3. Confusion matrix of different algorithms

3.3 The Evaluation Index

The credit card fraud detection model hopes to detect all fraudulent behaviors and not misjudge normal transaction behaviors. The author generated the confusion matrix under the above six combinations.

The confusion matrix consists of four values

Table 1. Confusion matrix

	Predicted value =1	Predicted value =0
True value =1	TP	FN
True value =0	FP	TN

TP: fraud was detected
FN: fraud was not detected
FP: normal trading practices were judged to be fraudulent
TN: normal trading behavior is judged to be normal
The following three values are introduced as indicators to evaluate the quality of the model.
Precision: (Frauds/ Detected Frauds)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

Recall:(Detected frauds/Actual frauds)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

F value:(A combination of precision and recall)

$$\text{F}_\beta = \frac{(1 + \beta^2) * \text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}} \tag{11}$$

β indicates the importance of precision. When β is greater than one , precision is more important than recall.

3.4 The Experimental Results

Conduct the experiments, and the Precision and Recall can be obtained according to confusion matrix, shown in table 2 and figure 4.

Table 2. Comparison results of different sampling methods

	Undersampling(Near Miss)			Oversampling (SMOPTE)		
	logistic	Neural Network	XGBoost	logistic	Neural Network	XGBoost
TN	56000	27877	55390	55000	56827	56840
FP	500	28989	1474	1600	37	24
FN	20	4	0	7	15	15
TP	78	94	98	91	83	83
PRECISION	0.135	0.003	0.062	0.053	0.692	0.776
RECALL	0.796	0.959	1	0.929	0.847	0.847
F (β=1)	0,231	0,006	0,117	0.102	0.761	0.810

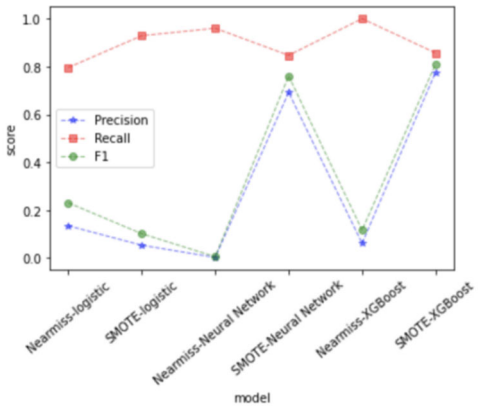


Figure 4. Comparison results

According to the above data, it can be found that Nearmiss-XGBoost has the highest Recall and SMOTE-XGBoost has the highest Precision. SMOTE-XGBoost has the highest value when β=1.

The index of Logistic is lower than other algorithms, indicating that neural network and XGBoost are more suitable for credit card fraud detection model. For these two algorithms, the Precision of NearMiss is generally low and the Recall value is high. This indicates that undersampling can well detect fraudulent behaviors, but the normal behaviors of customers are easy to be misjudged as fraudulent behaviors, which will affect the service quality of customers. The Precision is low because undersampling removes a large amount of normally behaving data. SMOTE Precision is high,and Recall is slightly lower. This means that it is possible to miss a very small number of frauds, but the probability of misjudgments is sharply reduced.

SMOTE sampling method is higher than NearMiss when F value β=1, which also explains that SMOTE is the main sampling method in most articles. However, omission of every fraud is likely to cause huge economic losses, and the value of recall is relatively important.

According to the actual situation, a credit card fraud detection model can be constructed to provide fraud likelihood score. For example, in this paper, the Nearmiss-XGBoost (Model A) and SMOTE-XGBoost(Model B) models are integrated. The fraud detected by model A is called set A, and the fraud detected by model B is called set B

$$A \cap B \text{ Five Scores}, \quad B - A \text{ Three Scores}, \quad A - B \text{ One Score}$$

In this way, five scores can be directly judged as fraud, while the other scores can pass different degrees of auxiliary detection. (use other factors, such as: comparison of historical records, etc.).

4. Conclusion

This paper introduces the basic construction of imbalanced data-credit card fraud detection model, and makes horizontal comparison between undersampling and oversampling, logistic regression, neural network and XGBoost algorithms.

According to the comparison, the results of the logistic regression algorithm are worse than those of the other two algorithms. SMOTE oversampling gives a good accuracy rate, while NearMiss undersampling gives a good recall rate. This indicates that NearMiss undersampling leads to some loss of majority class features by reducing majority class, SMOTE oversampling weakens minority class features by adding minority class, ensuring no wrong judgment on normal transaction behavior. In terms of algorithms, neural networks and XGBoost get better results than Logistic models. SMOTE-XGBoost gets the best overall result. It can be seen from the above results that XGBoost has better results than other algorithms under the condition of consent, and only the over-sampling and under-sampling synthesis can ensure that both precision and recall have relatively high values. In this paper, the basic architecture of credit card fraud model is introduced in simplified language, so that readers can understand the construction process and basic principles of the model. This is conducive to the face of different situations, different information, effective and effective plan. When different information has different importance, it can be optimized during standardization. Different sampling combination methods and algorithms are used according to the time cost. When it is necessary to provide a fraud level, consider the possibility of parallel comprehensive consideration of multiple models.

In the future, with the popularization of fraud detection model, new fraud means will inevitably appear. Faced with this situation, practitioners need to understand the rationale and infrastructure of the model. Starting from the demand, it provides a good optimization idea to ensure a faster change of the model, maintain a higher rate of credit card fraud detection, protect the interests of consumers, card owners and merchants, and maintain the order of the market.

References

[1] Sun Dan, Shi Weili, Rao Lanxiang, Meng Shasha, Guo Xiaoming, Li Yilun. Credit Card Fraud Detection Method Based on Improved Hybrid Sampling and XGBoost Algorithm [J]. Computer and Modernization,2022(09):111-118.

[2] Zhao Feng, Li Niuniu. Credit card fraud detection based on the mixed sample and the encoder application [J]. Journal of Harbin commercial university (natural science edition), 2022, 38 (4) : 420-426. The DOI: 10.19492 / j.carol carroll nki. 1672-0946.2022.04.005.

[3] Shi Xiangrong, Guo Pengsai, Zheng Qi, Ye Yifei. Application of ensemble Learning in Consumer Finance Audit: A case study of Random Forest in credit card fraud detection [J]. Business Accounting,2022(15):46-51.

[4] Hao Shipeng, Sun Yicheng. Change trend of credit card fraud risk and prevention and control suggestions [J]. China Credit Card,2022(03):41-45.

[5] Chen Rongrong, Zhan Guohua, Li Zhihua. Research on credit card transaction fraud prediction based on XGBoost algorithm model [J]. Application research of computers,2020,37(S1):111-112+115.

[6] Zhang Jiabei. Machine learning in the application of credit card fraud detection [D]. Central China normal university, 2021. The DOI: 10.27159 /, dc nki. Ghzsu. 2021.000520.

[7] Huang Yongxin. Based on the depth of the forest of credit card fraud detection research [D]. Jinan university, 2020. The DOI: 10.27167 /, dc nki. Gjinu. 2020.001135.

[8] Qi Shouxian, Hu Ronghui, Wang Wei, Wu Mengdi, Zhang Yuyong. Transformer Fault Diagnosis Based on SMOTE Balance Dataset [J]. Shandong Electric Power Technology, 222,49(04):15-22.

[9] Bao Jinghui, Yan Yu. The offense and defense of credit card transaction anti-fraud [J]. China Credit Card,2021(12):19-21.

[10] WU X Y. Research on credit card application anti-fraud scoring model of small and medium-sized banks in China [D]. Shanghai University of Finance and Economics,2021.