

Fraud detection based on FS-SMOTE model for credit card

Yiying Wang*

Nanjing University, Department of Computer Science and Technology, Nanjing, China, 210023,

*Corresponding author: yiyingwang82@gmail.com

Abstract. In the financial security technology, credit card fraud detection technology is an important technical means, which collects and analyzes the credit card transaction data in a certain period of time, detects fraud in many credit card transactions, and takes the corresponding alarm response. At the same time, in view of the extremely imbalanced characteristics of credit card fraud customer data set, the number of minority samples is increased by Smote, which is a representative algorithm of oversampling technology. Logistic regression, KNN, Decision tree, Bagging and Stochastic gradient descent are used to construct fraud detection models. The credit card fraud data released on Kaggle platform were selected for verification. The experimental results show that the fusion model based on Feature selection and Smote greatly improves the accuracy and detection rate of credit card fraud detection as a whole.

Keywords: Class-Imbalance, Oversampling, Smote, Feature selection, Fraud detection.

1. Introduction

Credit cards as a convenient payment tool, are widely used on a global scale. With the popularity of smartphones and the development of mobile payment technology, more and more people begin to use a mobile phone for credit card payment, increasing the convenience and security of the mobile payment, which will further promote the growth of credit card transactions. Credit card transactions have become one of the most popular payment methods both online and offline, and the occurrence of credit card fraud has increased with the rise in the number of credit card transactions. According to the Blue Book on the Development of China's Bank Card Industry (2022), in terms of transactions, the transaction amount of bank cards nationwide reached 1,060.6 trillion yuan, up 33.8% year-on-year. In terms of acceptance, the total number of merchants accepted in China was 27.983 million, down 3.3% year-on-year, while the total number of merchants accepted by UnionPay cards abroad exceeded 35 million, up 9.4% year-on-year. In terms of risks, the outstanding credit balance of bank cards was 8.62 trillion yuan, an increase of 8.9% over the previous year. The bank card fraud rate was 0.32BP, down 0.43BP from the previous year, and achieved five consecutive years of decline [1].

Now the common credit card fraud is:

- (1) Using forged and invalid credit cards to commit fraud;
- (2) Fraudulently using others' credit cards;
- (3) Malicious overdraft of credit card.

Credit card fraud is a kind of criminal deception for the purpose of obtaining economic benefits. China Banking Association mentioned in the Blue Book of China Bank Card Industry Development (2022) that in 2021, the development of bank card industry faced many challenges such as online fraud risk and external risk transmission. In this regard, the banking industry should adhere to the long-term and prudent risk management and constantly improve the digital anti-fraud capability. Therefore, the detection of credit card fraud has become one of the core capabilities of financial institutions.

Credit card fraud detection is to judge whether there is fraud by mining the information contained in the credit data of the cardholders, which is essentially a binary classification problem [2]. However, the data set of credit card fraud is extremely imbalanced, and the model cannot effectively mine the fraud sample information during training, which may lead to mistakes in the judgment. By processing imbalanced data, we hope to train efficient and stable models that can correctly identify fraudulent transactions.

2. Review of credit card fraud detection problems

The essence of credit card fraud detection is to classify a class of sequence data. Firstly, it constructed a binary classification detection model by processing the historical transaction data of credit cards, and then used the model to analyze the current behavior of cardholders and identify whether the behavior belongs to credit card fraud or not. In the traditional research on credit card fraud detection and credit measurement, Zhang Wen et al. used the improved decision support vector machine based on Kmeans to predict the risk of unbalanced P2P lending platform [3]. Jugovsky et al. firstly proposed to use Long Short-Term memory network (LSTM) model for fraud forensics in the financial field, which has better performance [4]. Srivastava et al. used the Hidden Markov Model (HMM) for the construction of fraud detection model, where they processed the sequence of operations in credit card transaction data [5]. Özçelik et al. used genetic algorithm to detect credit card fraud, which performed well in solving the problem of misclassification in credit card fraud detection [6].

3. FS-SMOTE credit card fraud detection model

3.1. Class-Imbalance

3.1.1 Definition

Class-Imbalance refers to the situation that the number of training samples of different classes varies greatly in classification tasks. When the training data is imbalanced, the number of rare instances in the data set is very small, which may cause these instances to be considered as noise and discarded during the algorithm learning. When dealing with the samples of fraud, because the number of positive and negative samples is extremely imbalanced, all the samples are classified as positive samples. Thus, even if the accuracy of the detection result is high, but no negative samples can be detected, in fact, the constructed model is meaningless. The goal of machine learning algorithm is to be as accurate as possible, which means that imbalanced datasets should be treated first and then try to improve the accuracy of the model. This experiment uses Smote algorithm.

3.1.2 Smote

The synthetic minority oversampling technique (Smote) algorithm is an improvement over random oversampling based on "linear interpolation" to synthesize new samples for minority classes [7]. For each minority sample X_i , select a minority sample $\hat{x}_i$ at random from its nearest neighbor, then randomly select the points on the line between X_i and $\hat{x}_i$ as the newly synthesized minority sample. A few newly synthesized class samples are added to the original dataset, so the generated samples have a characteristic attribute correlation with the samples in the original dataset [8]. The new data set can be used as a training sample to train the classification model. Since simple random oversampling is the direct reuse of a small number of class samples, the newly generated training set has many repeated samples, which will cause overfitting of the model. Smote algorithm fundamentally solves this problem (Table 1).

Table 1. Algorithm 1: Smote

Algorithm 1: Smote	
Input: set T , oversampling rate N , KNN parameter k , minority class sample size n .	
Output: new training set S	
1.	Compute k neighbors of small-sized sample x_i
2.	Choose a random sample x_a from the k nearest neighbors
3.	Generate a random number δ between 0 and 1 to synthesize a new sample x_{i1} , $x_{i1} = x_i + \delta * (x_a - x_i)$
4.	Repeat Step 2-3 at the oversampling rate N
5.	Get a new training sample S
6.	Classify the new training samples S by a classifier
7.	Output classification result

3.2. Feature selection

In ensemble learning, feature selection is mainly to filter the features of the data set, construct the feature subset of the data set without losing the information contained in the data, reasonably compress the feature space, and finally reduce the difficulty of model learning. To avoid overfitting, decrease the number of features to enhance the generalization ability of the model. The forms of studying feature selection can be divided into three types: Filter, Wrapper, and Embedded [9].

Filter is based on the general performance of features to select, independent of the model algorithm used, from the single variable itself and the relationship between the multivariate variable to determine whether the variable should be filtered out. Embedded takes the feature selection algorithm as a sub-process of the algorithm training process. Wrapper evaluates the effect of feature selection by using classifier performance indicators, its feature selection strategy needs to be changed when the model is changed.

The process of feature selection is as follows: suppose the original data set has n dimensional features, and the feature vector is $X = \{x_1, x_2, \dots, x_n\}^T$, after feature selection, the feature is reduced to m dimensions, and its feature vector changes to $Y = \{x_1, x_2, \dots, x_m\}^T$, where $n > m$.

3.3. FS-SMOTE credit card fraud detection model building

In FS-SMOTE credit card fraud detection model, select features of the data set and divide the data set. The imbalanced training set is processed through the Smote algorithm and the undersampling algorithm so that the minority class samples can be generated. Then three classifiers are selected for classification, and the voting classifier is used to integrate the three results to improve the detection performance of the model. Figure 1 shows the process of model construction and implementation (Fig.1).

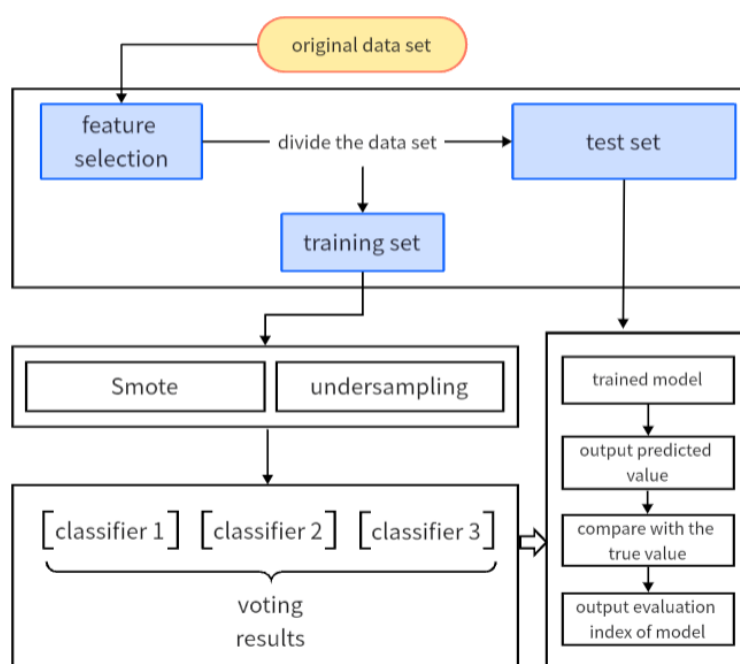


Fig. 1 FS-SMOTE credit card fraud detection model

4. Empirical research

4.1. Data sources

The dataset used in this paper is credit card fraud data published on Kaggle. The dataset was collected by Worldline and machine learning group at Université libre de Bruxelles. It contains information on transactions that occurred over two days in September 2013 for selected European

credit card holders. The dataset consists of 284,807 transactions with credit cards, of which 492 were fraudulent. The data contains a total of 30 features and a label. In the label category, 0 indicates normal trading and 1 indicates fraudulent trading.

4.2. Data preprocessing

In this paper, due to the large difference between the value range of the amount column and time column, the data of the amount column and time column is normalized, and the rule of normalization is shown in the formula (1).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

After normalization of the dataset, the number of features remained is 30. Do feature selection on the processed data set.

4.3. Feature selection

The basic idea of single variable selection method is to choose variable with the highest correlation with the dependent variable and discard variables that are not related to the target variable. The F-test or chi-square test is used to calculate the correlation with the dependent variable. Since this is a classification problem, we use python's `f_classif` function to compute the score. The higher the score, the more strongly the feature is associated with the target variable.

The features are scored using the `SelectKBest` and `f_classif` methods, and the `k` features with the highest scores are selected (Fig.2).

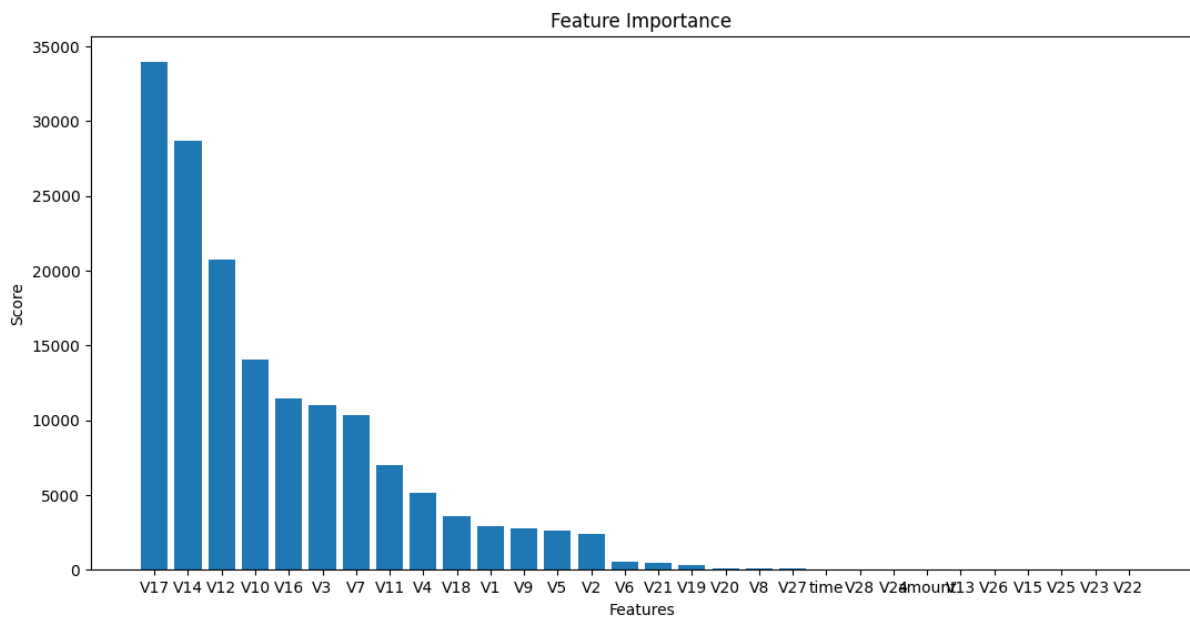


Fig. 2 Importance of features

Based on the importance of features ranking from high to low, 14 to 30 features are to be selected.

4.4. Imbalanced sample processing

The credit card transaction information used in this experiment is imbalanced data, the proportion of fraud samples in the original data set is only 0.173%, while the proportion of legitimate samples is 99.827%, the proportion of fraud samples and legitimate samples is extremely different. Considering that the given data is highly imbalanced, the minority class samples may be regarded as noise and discarded during model learning. In order to reduce the deviation between data and the loss caused by misclassification by the machine learning algorithm, it is necessary to balance the data in some way so that the model can be effectively trained.

Considering that there may be discrepancies between the prediction results of the model on the test set data after processing of imbalanced data and the prediction results of the original data set, this paper first divides the preprocessed data set, and then uses the oversampling representative algorithm Smote for the training set, and synthesizes the new samples of the relatively few fraud samples to increase the sample size. After synthesis, the legal samples accounts for 2/3 of the new training set, and the fraud samples accounts for 1/3 of the new training set, thus completing the handling of imbalanced data. Table 2 shows the proportion of positive and negative samples in the training set.

Table 2. Training set and test set after processing

Data set	Fraud sample	Legal sample	Total
Original data set	492	283415	284807
Proportion	0.173%	99.827%	1
Training set	21323	42646	63969
Proportion	33.3%	66.7%	1

4.5. Model training

Classifiers such as Logistic regression, KNN, Decision Tree, Bagging, and Stochastic gradient descent are used for model training, and the test set is used to evaluate performance of the models. Model parameters are shown in Table 2.

Table 3. Model parameters

Model	Parameters
Logistic Regression	penalty: "l2", dual: False, tol: 1e-4, C: 1.0, fit_intercept: True, intercept_scaling: 1, class_weight: None, solver: "lbfgs", max_iter: 100, multi_class: "auto", verbose: 0, warm_start: False, n_jobs: None, l1_ratio: None
KNN	n_neighbors: 5, weights: "uniform", algorithm: "auto", leaf_size: 30, p: 2, metric: "minkowski", metric_params: None, n_jobs: None
Decision Tree	criterion: "gini", splitter: "best", max_depth: None, min_samples_split: 2, min_samples_leaf: 1, min_weight_fraction_leaf: 0.0, max_features: None, random_state: None, max_leaf_nodes: None, min_impurity_decrease: 0.0, class_weight: None, ccp_alpha: 0.0
Bagging	estimator: None, n_estimators: 10, max_samples: 1.0, max_features: 1.0, bootstrap: True, bootstrap_features: False, oob_score: False, warm_start: False, n_jobs: None, random_state: None, verbose: 0, base_estimator: "deprecated"
SGD	loss: "hinge", penalty: "l2", alpha: 0.0001, l1_ratio: 0.15, fit_intercept: True, max_iter: 1000, tol: 1e-3, shuffle: True, verbose: 0, epsilon: DEFAULT_EPSILON, n_jobs: None, random_state: None, learning_rate: "optimal", eta0: 0.0, power_t: 0.5, early_stopping: False, validation_fraction: 0.1, n_iter_no_change: 5, class_weight: None, warm_start: False, average: False

4.6. Result analysis

This section from two parts, evaluation index and performance comparison, to introduce assessment methods and compare results of detection.

4.6.1 Evaluation metric

Table 4. Confusion matrix

	Predicted to be fraudulent	Predicted to be legal
Actually fraudulent	TP (True Positive)	FN (False Negative)
Actually legal	FP (False Positive)	TN (True Negative)

The existing evaluation metrics for binary classification first evaluate all instances in the project based on the confusion matrix of the above four cases, see Table 4.

In this paper, accuracy, precision, recall, F1_score, and AUC_ROC are used as model evaluation indicators. $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, $F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$, $Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$. With FPR the horizontal axis, TPR the longitudinal axis, map the ROC curves. The performance of the model is evaluated by the range covered by the ROC curve. The AUC value is the area under the ROC curve and is used to compare performance of different models. The higher the value, the better the performance of the model.

4.6.2 Performance comparison

In this paper, Logistic regression model is compared with KNN, Decision tree, Bagging and Stochastic gradient descent model. Fraud detection results of each model are shown in Figure 3 and Table 5.

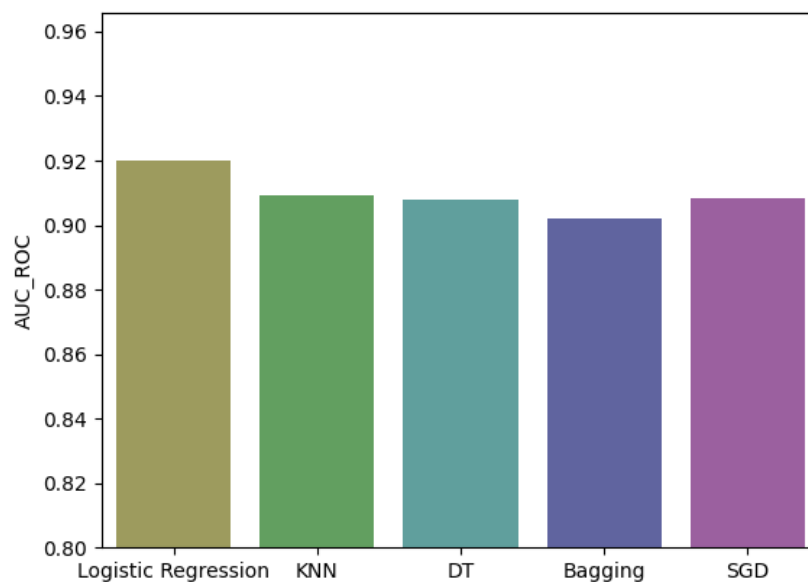


Fig. 3 Fraud detection results of each model

Table 5. Fraud detection results of each model

Model	AUC_ROC
Logistic Regression	0.9200
KNN	0.9094
Decision Tree	0.9078
Bagging	0.9021
SGD	0.9084

It can be seen from Table 5 that the prediction results of three models, KNN, Decision Tree and SGD, are very close, and their prediction results are slightly worse than the more effective Logistic model. Bagging model shows worse comprehensive classification ability than the above four models. The possible reason of Logistic model getting better evaluation results is that the generalization ability of the model is enhanced after the introduction of l2 penalty term.

In this paper, the classification results of the three classifiers are integrated by voting. First of all, choose three classifiers (Logistic Regression, KNN, SGD) with better performance. From there, we do feature selection again and compare the performance of fraud detection when choosing 14-30 features. Results are shown in Figure 4 and Table 6.

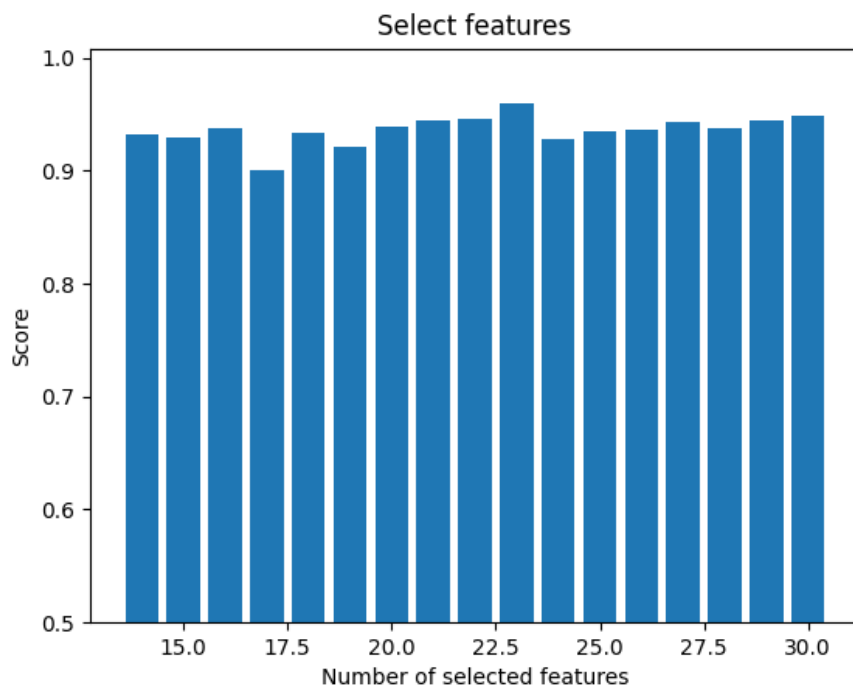


Fig. 4 Fraud detection results with different number of selected features

Table 6. Fraud detection results with different number of selected features

Number of selected features	Score
23	0.9604
30	0.9486
22	0.9453
29	0.9447
21	0.9443
27	0.9426
20	0.9392
16	0.9381
28	0.9371
26	0.9364
25	0.9349
18	0.9340
14	0.9321
15	0.9301
24	0.9282
19	0.9216
17	0.9010

The fraud detection model performs best when the number of selected features is 23. It has the highest AUC_ROC value of 0.9604 and the accuracy is as high as 98.57%. At this point, the confusion matrix and the values of each evaluation index are shown in Table 7&8.

Table 7. Confusion matrix when the number of selected features = 23

Actual Fraud \ Flagged Fraud	0	1
0	70065	1001
1	17	119

Table 8. Values of evaluation index when the number of selected features = 23

Evaluation index				
Accuracy	Precision	Recall	F1-score	AUC_ROC
0.985763	0.9997	0.985964	0.992831	0.9604585

It shows again that FS-SMOTE model has strong ability of generalization and classification.

5. Conclusion

The experimental results show that the fusion model based on Feature selection and Smote algorithm greatly improves the detection rate of credit card fraud detection as a whole. If the features of the sample are more, the efficiency of the model for fraud detection will be lower. The problem of extremely imbalanced sample is solved as well by Smote algorithm, it can influence the accuracy of the model to detect fraud.

In this paper, deep learning technology is also used to establish a model to detect credit card fraud. The results show that the accuracy of the model is 98.57%, AUC_ROC value is 0.9604, indicating that FS-SMOTE model can deal with the imbalanced data in credit fraud well and provides some theoretical basis and practical reference for financial institutions to detect the risk of financial fraud.

References

- [1] China Banking Association. China's bank card industry development blue book (2020) [EB/OL]. (2020-09-24) [2021-09-18]. http://jrj.wuhan.gov.cn/ztlz_57/xyrd/yxy/202009/t20200924_1455007.sht ml.
- [2] Liu Bo, Liang Longyue. Credit card fraud detection based on KM - SVM SMOTE - CNN [J]. Computer system application, 2022, 31 (6): 361-367. The DOI: 10.15888/j.carol carroll nki. Csa. 008536.
- [3] Zhang W, Cui Y B, Jiang Y P. A study on risk prediction on unbalanced P2P lending data based on SVM K-Mean. Journal of Systems Science and Mathematical Sciences, 2018, 38(3): 364-378.
- [4] Jurgovsky J, G ranitzer M, Ziegler K, et al. Sequence classification for credit card fraud detection. Expert Systems with Applications, 2018, 100, 234-245.
- [5] Srivastava A, Kundu A, Sural S, et al. Credit card fraud detection using hidden Markov model. IEEE Transactions on Dependable and Secure Computing, 2008, 5(1): 37-48. [doi: 10.1109/TDSC.2007.70228]
- [6] Özçelik MH, Duman E, Işık M, et al. Improving a credit card fraud detection system using genetic algorithm. 2010 International Conference on Networking and Information Technology. Manila: IEEE, 2010. 436-440.
- [7] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.
- [8] Ju C H, Bao F G, Xu C H, et al. A novel method of interestingness measures for association rules mining based on profit. Discrete Dynamics in Nature and Society, 2015, DOI: 10.1155/2015/868634.
- [9] Chen B, Su Y D, Huang S. Classification of unbalanced data based on KM-SMOTE and Random Forest [J]. Computer Technology and Development, 2015, 25(09): 17-21.