IMPACT: International Journal of Research in Engineering and Technology
ISSN (P): 2347–4599; ISSN (E): 2321–8843
Vol. 10, Issue 4, Apr 2022, 1–10

ANALYSIS AND RESEARCH ON CREDIT CARD FRAUD DETECTION BASED ON MACHINE LEARNING

Linyu Wei[1], Wanting Chen[2], Shuwen Chen[3], Xuanyi Wu[4], Yahui Meng[5], ZY Chen[6], Ruei-yuan Wang[7] & Timothy Chen[8]

[1,2,3,4,5,6,7]Research Scholar, School of Science, Guangdong University of Petrochemical Technology, Maoming, Guangdong 525000, China

[8]Research Scholar, Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA 91125

Received: 01 Apr 2022 **Accepted:** 04 Apr 2022 **Published:** 05 Apr 2022

ABSTRACT

With the rapid development of socialist economy, there are many financial payment tools in our life, such as various credit cards issued by financial institutions. Although its emergence has brought us a lot of convenience, but at the same time there are certain disadvantages. In recent years, the crime of credit card swindler is increasing year by year. In view of whether there is a fraud in the credit card, this paper uses SMOTE algorithm to balance the unbalanced data. Again using Logistic regression analysis algorithm, XGBoost algorithm respectively, KNN algorithm with UCI data set Default of credit card clients data analysis and research, found that under the unbalanced data sets of different accuracy of classification algorithms, among them, XGBoost algorithm and KNN algorithm have higher accuracy results and are suitable for the classification of this data set.

KEYWORDS: *Credit Card Fraud*

INTRODUCTION

Credit card fraud, mainly from credit card risk. [1] the credit card fraud risks with its bad social impact, brings enormous economic losses, and growing as the Internet technology, all kinds of system increasingly complex for its good concealment, credit card fraud in the present moment has a wider stage, and gradually became the restraints on the development of Banks and card issuers of credit card industry's leading adverse factors, It becomes an urgent problem to be solved by relevant institutions and scholars. Therefore, this paper uses machine learning [2] to analyze and study credit card fraud detection.

FRAUD DETECTION

Algorithm Theory

SMOTE Algorithm

SMOTE data balance. [3] Synthetic minority class oversampling technology, which is an improved scheme based on random oversampling algorithm, because random oversampling adopts the strategy of simply copying samples to increase minority class samples, it is easy to produce the problem of model over fitting, even if the information learned from the model is too Specific rather than General. The basic idea of SMOTE is to analyse a few sample and add a new one to the set. Its calculation principle is shown in Formula 1.

$$x_{new} = x + rand\ (0,1) \times (\tilde{x} - x) \quad (1)$$

Logistic Regression Analysis

Logistic regression analysis. [4-6] Also known as logistic regression. Logistic regression analysis method is a kind of nonlinear regression to analyze the probability of the independent classification data type statistical method, applicable to the dependent variable as classification variables (binary classification or multiple classification), the independent variables or continuous variables for classification, and the variance of data and normality does not make specific requirements, and therefore is widely used in different fields [5]. This algorithm can be used for both classification and regression, but it is widely used for classification. The advantage is that the calculation cost is not high, easy to understand and implement. The disadvantage is that it is easy to fit and the classification accuracy may not be high. Applicable data types: numerical and nominal data. Its calculation principle is shown in Formula 2.

$$p = 1/1 + e(a+ax+ax+\cdots+01122....anxn) \quad (2)$$

XGBoost Algorithm

XGBoost is a tree integration model that combines multiple weak classifiers into one strong classifier [7]. XGBoost combines the prediction results of multiple trees, improves the generalization ability of the model, can automatically learn the processing strategy of missing values, and adopts the strategy similar to random forest to sample data, aiming to achieve high efficiency, flexibility and portability. [8] Make the predicted results more accurate. Its advantages are higher precision, more flexibility, parallel approximation algorithm, which can be used to generate candidate segmentation points. Its calculation principle is shown in Formula 3 [9]

$$\tilde{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (3)$$

Knn Algorithm

KNN algorithm is a simple and classical machine learning classification method. The samples are classified by measuring the distance (usually using Euclidean distance) or similarity between the samples to be classified and the samples of known categories [10]. KNN classification algorithm indicates that any sample is represented by k neighbors closest to it. According to the distance distribution between the training sample and the nearest neighbor sample, the threshold with a certain confidence level is determined [11]. Is a more mature method in theory simple and easy to use, easy to understand, high precision, mature theory. [12]

Confusion Matrix

The obfuscation matrix provides more knowledge about the performance of our model, it provides information about correct and incorrect classification, and it allows us to identify errors. So that the information is more accurate. Figure 1 shows how the confusion matrix is formed.

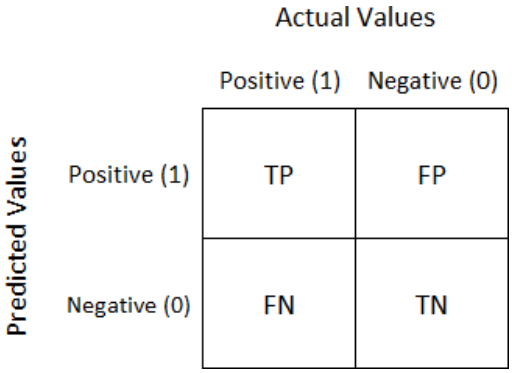

Figure1: Formation of Confusion Matrices.

The Experiment Design

Any kind of fraud is harmful to the society, and with the increasing development of science and technology, there are more and more credit card fraud, but there are still many difficulties in the study of credit card fraud. This study compares the accuracy of six data mining methods in predicting the probability of delinquency for the case of customers in Taiwan. From a risk management point of view, the result of estimating the predictive accuracy of the probability of default will be more valuable than the binary result of categorization -- trusted or untrusted customers. Since the true probability of default is unknown, this study proposes a novel "rank smoothing method" to estimate the true probability of default. Taking the actual probability of default as the response variable (Y) and the probability of default prediction as the independent variable (X), the simple linear regression results (Y=A+BX) show that the prediction model generated by artificial neural network has the highest determination coefficient. The regression intercept (A) is close to zero, and the regression coefficient (B) is close to 1. Therefore, among the six data mining technologies, artificial neural network is the only one that can accurately estimate the real probability of default.

Figure 2 shows the method and design of this experiment.

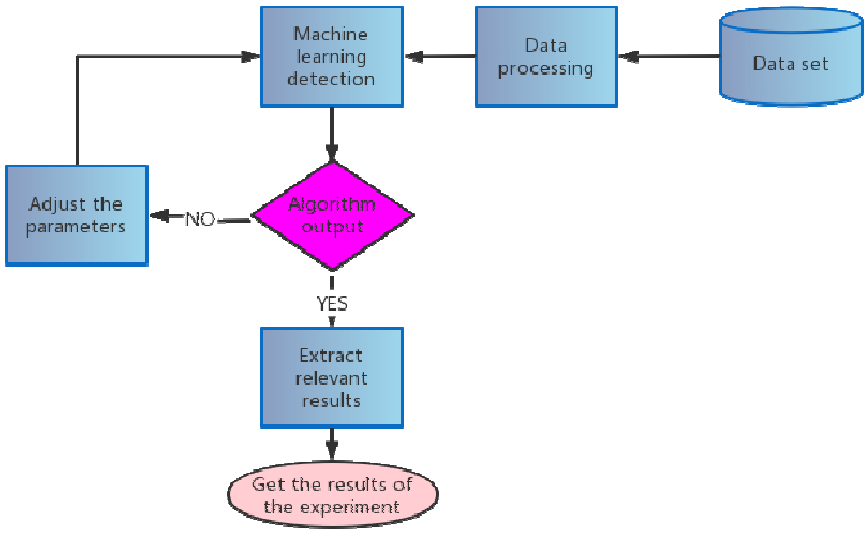

Figure 2: Experimental Design.

The Data Processing

Variable Selection

The binary variable default payment (yes = 1, no = 0) was used as the response variable in this study. This study reviewed the literature and used the following 23 variables as explanatory variables, as shown in Table 1:

Table 1: Explanatory Variables

Characteristic	Content	Annotation	Data Sources
X1	Given Credit Limit (NTD)	Includes personal consumption credits and his/her family (supplementary) credits.	UCI dataset
X2	sex	1 = male; 2 = female	UCI dataset
X3	education	1 = Graduate School; 2 = University; 3 = High School; 4 = Other	UCI dataset
X4	marital status	1 = Married; 2 = Single; 3 = Other	UCI dataset
X5	Age (year)		UCI dataset
X6 - X11	History of past payments. We track past monthly repayments (April-September 2005) as follows:	X6 = Repayments in September 2005; X7 = Repayments in August 2005; X11 = Repayments in April 2005. The repayment is measured by: -1 = repayment on time; 1 = payment delay of one month; 2 = payment delay of two months; . . . . ; 8 = Payment delay of eight months; 9 = Payment delay of nine months or more.	UCI dataset
X12-X17	Bill amount (NTD).	X 12 = September 2005 statement amount; X 13 = August 2005 statement amount;. ..; X 17 = Bill amount for April 2005.	UCI dataset
X18-X23	The amount of the last payment (NTD).	X 18 = Amount paid in September 2005; X 19 = amount paid in August 2005;. ..; X 23 = amount paid in April 2005.	UCI dataset

The Data Collection

The data set was selected from the Default of Credit Card Clients data set of UCI. This data set has 25 columns, 23 of which are the characteristics of the data set (X1 to X23), and the column variable Y is the category label. Since the dimensionality of each feature of the data is different, the data set is first standardized using Standard Scaler (), as shown in Figure 3, and variable "ID" and class-related data are eliminated during the standardization.

```
array([[-1.13672015,  0.81016074,  0.18582826, ..., -0.30806256,
        -0.31413612, -0.29338206],
       [-0.3659805 ,  0.81016074,  0.18582826, ..., -0.24422965,
        -0.31413612, -0.18087821],
       [-0.59720239,  0.81016074,  0.18582826, ..., -0.24422965,
        -0.24868274, -0.01212243],
       ...,
       [-1.05964618, -1.23432296,  0.18582826, ..., -0.03996431,
        -0.18322937, -0.11900109],
       [-0.67427636, -1.23432296,  1.45111372, ..., -0.18512036,
         3.15253642, -0.19190359],
       [-0.90549825, -1.23432296,  0.18582826, ..., -0.24422965,
        -0.24868274, -0.23713013]])
```

Figure 3: Standardized Data.

Data Standardization

The new standardized data set data X was obtained after standardization, and the correlation visualization test was carried out on the standardized 23 variable data, and the correlation between data was further analyzed by drawing the thermal diagram of relational coefficients. As shown in FIG. 4, it can be seen from the output thermal diagram that there are very conspicuous dark parts along the way, indicating that there is an obvious local linear correlation between variables in this data set, such as strong linear correlation between X6-X11 and X12-X17.

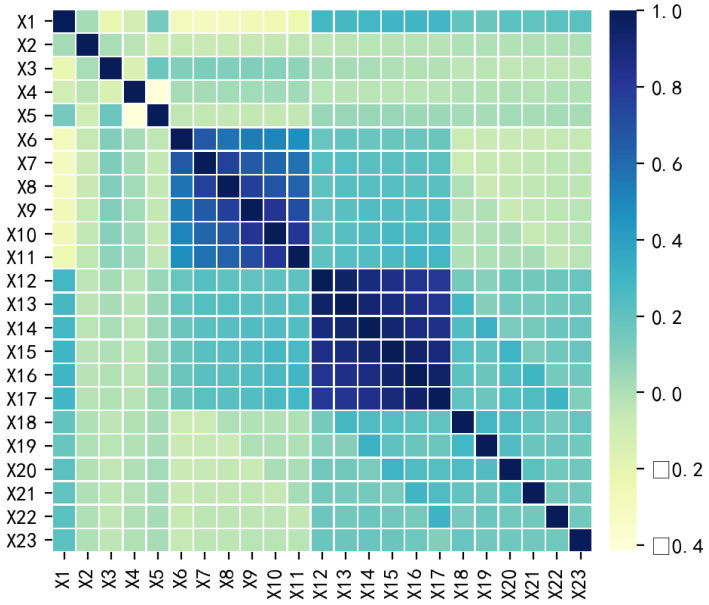

Figure 3: Correlation Analysis Plot of 15 Principal Components (No Dimensionality Reduction).

Data Dimension Reduction

Direct use of original data for modeling may affect the stability of the model. To solve this problem, principal component analysis can be used to reduce the dimension of data and extract its main components for the establishment of the model. The fit_transform method of PCA () was used to reduce the dimensionality of the standardized data set, and the first 1 and the first principal component were selected to retain 95.7% of the information in the original data.

Data Correlation Processing

After dimensionality reduction of the data set by principal component analysis, there is no linear relationship between the 15 principal components. In order to increase the accuracy of the classification model, we first drew a thermal diagram to test the linear correlation between the 15 principal components, as shown in FIG. 5. Before establishing the classification model, we need to judge whether the sample data of classification variables are balanced or not. A serious imbalanced sample data will lead to a large probability that the trained model will output the category with a large number, which will make the model have a strong bias, thus reducing the accuracy of classification. In order to improve the accuracy of classification model. It can be seen from FIG. 6 that the number of samples in this data set is not balanced. 数据降维.

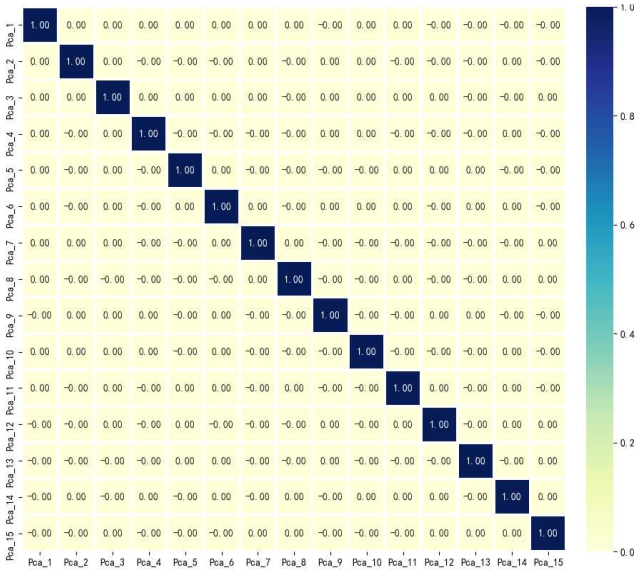

Figure 4: Correlation Analysis Plot of 15 Principal Components (Dimensionality Reduced).

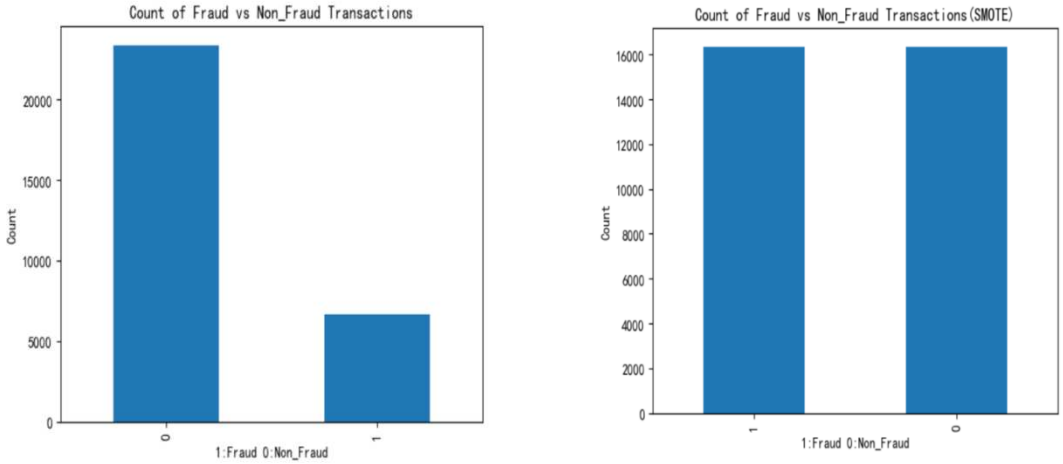

Figure 5: Fraud Classification Histogram.

Data Balancing

When dealing with the problem of category imbalance, it can only be processed on the training set, not on the test set, so it is necessary to slice the data set first. 20% of the 30,000 solar calendars were taken as the test set to perform data cutting, as shown in Table 2. As can be seen from Table 2, the data instances in the training set are seriously unbalanced. If the unbalanced data sets are directly used for classification training, the results of the model will be closer to the large category, resulting in high accuracy. In order to make the model more accurate, we use SMOTE over-sampling to balance the data.

Balance the training data using FIT_sample () in SMOTE. According to the result, after the balance, the sample number of 1 and 0 is 18,661, and the class with few original cases has been sampled, as shown in Table 3.

Table 2: Cut the Training Dataset by Categorical Variables

	0	1
Data	18661	5339

Table 3: Over-Sampling Equilibrium

The number of data collected by SMTO	0	1
Training Set	18661	18661
Test Set	4703	4703

Experiments and Results

Here, this paper uses Logistic regression analysis, Xgboost algorithm, KNN algorithm to classify the credit card fraud detection, also use SMOTE data balance and confusion matrix to judge the prediction error, by the accuracy and recall rate calculated by each algorithm to find the best algorithm.

Classification of Credit Card Fraud Detection Based on Logistic Regression Analysis

The experiment uses Python to analyze the default of credit card clients in THE UCI dataset by using the existing SKLearn machine learning library, and then classifies them by Logistic regression analysis to further calculate the predicted values of feature vectors. Finally, the confusion matrix is visualized to judge the prediction error. The specific results are shown in Table 4 and Figure 7. The accuracy calculated by this method is 74.13%, but the recall rate is only 70.12%. In the confusion matrix, the position 00 represents 5565 bank cards without fraudulent behavior, and the model predicts that 5565 bank cards have no fraudulent behavior. Position 01 represents 1446 bank cards without fraudulent behaviors, while the model predicts that 1146 bank cards have fraudulent behaviors. The position 10 represents that there are 1,832 cards that are not fraudulent, but the model predicts that 1,832 cards are fraudulent. The position 11 represents that 4145 bank cards are actually fraudulent, and the model also predicts that 4145 bank cards are fraudulent. It can be seen from the visual confusion matrix that Logistic accuracy is not high enough and recall rate is relatively low, so it is not suitable for the classification calculation of this data set.

Table 4: Logistic Classification Results

AUC	0.72248324062188
ACC	0.72248324062188
Recall	0.7012138068749109
F1-score	0.7378320901444214
Precesion	0.7413700590234306

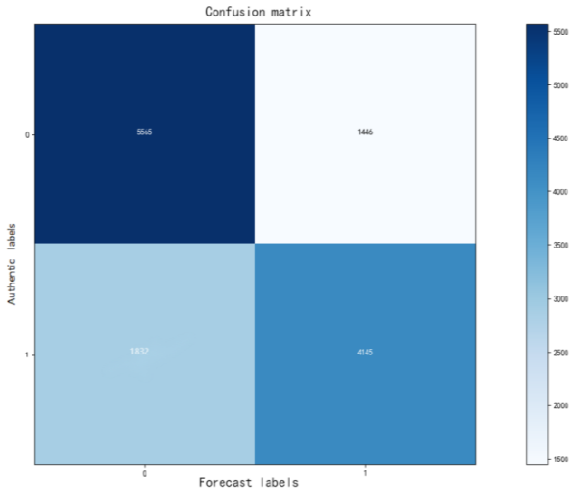

Figure 6: Logistic Regression Analysis Confuse Matrices.

Credit Card Fraud Detection is Classified Based on XGBoost Algorithm

The default of credit card clients in THE UCI dataset is analyzed using the Existing SKLearn machine learning library in Python. XGBoost algorithm is used to classify the default of credit card clients, and the predicted values of feature vectors are further calculated. Finally, the confusion matrix is visualized to judge the prediction error. The results are shown in Table 5 and Figure 8.

Table 5 shows that the accuracy of the optimized tree integration model is up to 85.76%, and the regression rate is further improved to 75%. Due to the limited amount of data in the Default of Credit Card Clients data set, there is no way to continue to optimize the accuracy due to more iterations. Figure 8 shows that in the confusion matrix, the position 00 represents 4046 bank cards without fraudulent behaviors, and the model predicts that 4046 bank cards have no fraudulent behaviors. Position 01 represents that 562 bank cards have no fraudulent behaviors, while the model predicts that 562 bank cards have fraudulent behaviors. Position 10 represents that 1161 bank cards have no fraudulent behaviors, but the model predicts that 1161 bank cards have fraudulent behaviors. The position 11 represents 3505 bank cards with real fraudulent behavior, and the model also predicts 3505 bank cards with fraudulent behavior. It can be seen from the visual confusion matrix that XGBoost algorithm has significantly higher accuracy and recall rate than Logistic algorithm, which is suitable for the classification calculation of this data set.

Table 5: XGBoost Algorithm Classification Results

	train-logloss	test-logloss
0	0.54775	0.59037
1	0.45987	0.53859
2	0.39666	0.50409
3	0.34791	0.47969
4	0.30811	0.46146
5	0.27856	0.45149
6	0.25317	0.44392
7	0.22936	0.42992
8	0.20864	0.41759
9	0.19367	0.41696

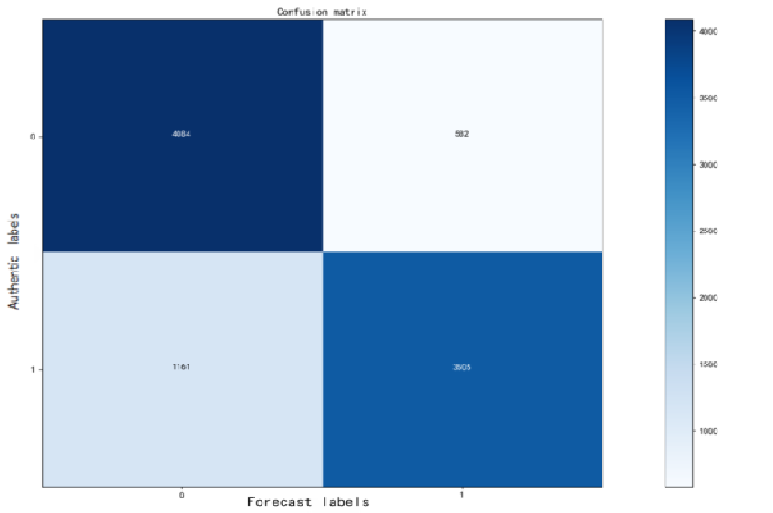

Figure 7: XGBoost Algorithm Classification Confusion Matrices.

This Paper Classifies Credit Card Fraud Detection Based on KNN Algorithm

The experiment was implemented using Python, and the existing SKlearn machine learning library was used to analyze the default of Credit card Clients in the UCI dataset, so as to further calculate the predicted value and accuracy of feature vectors. The specific results are shown in Table 6 and Figure 9.

Table 6 shows that the accuracy of the optimized tree integration model is up to 82%, while the regression rate is further improved to 98.6%. Due to the limited amount of data in the Default of Credit Card Clients data set, there is no way to continue to optimize the accuracy due to more iterations. Figure 9 shows that in the confusion matrix, the position 00 represents 18,660 bank cards without fraud, and the model predicts that 18,660 bank cards have no fraud. The position of 01 indicates that there is no fraudulent behavior in one bank card, while the model predicts that there is fraudulent behavior in one bank card. Position 10 represents that there are 1530 bank cards with no fraudulent behavior, but the model predicts that 1530 bank cards have fraudulent behavior. The position 11 represents that 17,131 bank cards are actually fraudulent, and the model also predicts that 17,131 bank cards are fraudulent. It can be seen from the visual confusion matrix that the ACCURACY and recall rate of KNN algorithm are obviously high, which is suitable for the classification calculation of this data set.

Table 6: KNN Classification Results

	V1	V2	V3	V4
Accuracy			0.82	37322
Macroavg	0.83	0.82	0.82	37322
Weighted ayg	0.83	0.82	0.82	37322

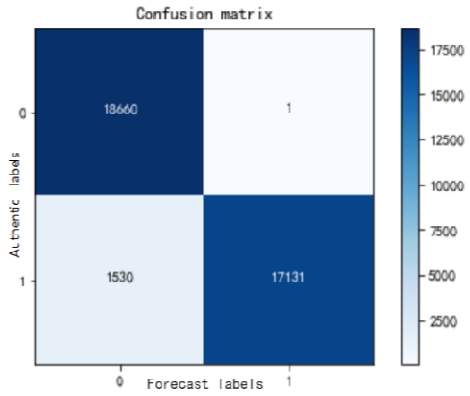

Figure 8: Confusion Matrix Under the KNN Classification.

CONCLUSION

This paper analyzes and studies the Default of Credit card Clients data in the UCI dataset and finds that the accuracy of different classification algorithms varies under the unbalanced dataset, as shown in Table 7.

Table 7: Accuracy of the Different Algorithms

	Accuracy	Recall	Suitable for Unbalanced Dataset Classification
Logistic Regression Analysis Algorithm	74.13%	70.12%	NO
XGBoost Algorithm	85.76%	75%	YES
KNN Algorithm	82%	98.6%	YES

Finally, this paper studied the Logistic regression analysis, XGBoost algorithm, Knn algorithm and other machine learning algorithms. The results show that XGBoost algorithm has a higher accuracy of 85.76%. The recall rate of KNN algorithm was 98.6%. Because this paper uses many algorithms, but did not reach 100% accuracy. But we're also trying to get better accuracy. Even higher results can be achieved if more data can be obtained for experiments and computer algorithms are used, which will be tested in future work.

REFERENCE

1. *Lin Wanxuan. Research on credit card risk Prevention [J]. Shenzhou, 2013(21):202.*
2. *Li Haoquan, Shi Mengfan, Chen Shonan, Zhang Junyang. Application of convolutional neural network in case classification [J]. Software, 2019,40(04):222-225.*
3. *SRIPT's Novel SMTO Catalyst Has Passed Appraisal[J].China Petroleum Processing & Petrochemical Technology, 2021, 23 (4) : 74.*
4. *Wang Yang, XU Gangfeng, Zuo Dongguang. Infrared Technology, 2008(10):563-566. (in Chinese)*
5. *Li Jiaxin. Analysis of personal credit Evaluation based on Stepwise Logistic Regression Classification Algorithm [J]. Journal of Hunan University of Arts and Sciences (Natural Science Edition), 201, 33(01):5-8+57.*
6. *Xu Yiping, Zhuang Lingxi. International Comparative Study on industrial policy of 5G: Based on dichotomous Logistic regression Analysis [J]. Scientific management, 2021, 33 (05) 6:87-94. The DOI: 10.19445 / j.carol carroll nki/g3.2021.05.013. 15-1103.*
7. *Wang Xiaoyi, Wang Ziyi, Zhao Zhiyao, Zhang Xin, Chen Qian, Li Fei. Integrated improved AHP and XGBoost algorithm for food safety risk prediction model: A case study of rice [J]. Journal of food science and technology: 1-9 [2022-03-06]. HTTP: / / http://kns.cnki.net/kcms/detail/10.1151.TS.20220223.1026.008.html.*
8. *Zhang Xingzhi. Improving SMOTE method based on XGBoost credit score model [J]. Network Security Technology and Application,2022(02):37-41.*
9. *Zhu Yuanqing, Li Saifei, Li Hongzhe. Host attack detection based on XGBoost and community found [J]. Computer system application, 2021, 30 (12) : 147-154. The DOI: 10.15888 / j.carol carroll nki. Csa. 008211.*
10. *Wang Xiaoyun, WANG Dongqin, GUO Jinyu. Multiple modal based on markov distance kNN process fault detection [J]. Journal of shenyang university (natural science edition), 2021 (6) : 480-485. The DOI: 10.16103 / j.carol carroll nki. 21-1583 / n. 2021.06.004.*
11. *Xie Miao, Lin Yongchang, Zhu Xiaoshu. Journal of Hefei University of Technology (Natural Science), 201,44(11):1483-1486+1505.*
12. *Piao Hongde, Jinshanhai. Bad Gait Classification Based on KNN Algorithm [J]. Information Technology and Informatization,2022(01):190-193.*