Credit Card Fraud Detection Using Predictive Model

Jiayin Zhang *

Department of Mathematics, University of California, Santa Barbara, Santa Barbara, United States

*Corresponding author: jiayin_zhang@ucsb.edu

Abstract. Credit card is a sign of credit that is given to customers with good credit by a commercial bank or credit card firm. It takes the shape of a card with signature blank space on the back and the name of the dissipated bank, expiration date, CVS number, and cardholder name on the front. A credit card is a payment card that can give the cardholders' abilities to enable the cardholder to exchange for goods and services based on their credibility and debt score. In this paper, it will explore the credit card fraud detection predictive model to avoid fraudulent activity. Through different algorithms, the study could easily show the potentially fraudulent activities in the given dataset. In order to effectively combat credit card fraud, a number of techniques have been developed and put into practice, including different supervised and unsupervised machine learning algorithms to predict fraudulent activities. These techniques will be used to compare between the actual dataset and estimated models to illustrate the full picture. The credit card fraud is a challenging problem, especially it is prevalent during college students.

Keywords: Machine Learning Algorithms; Predictive Model; Fraud detection.

1. Introduction

With the rapid expansion of the credit card business and the increasingly fierce competition, all kinds of fraud risks have become the biggest threat in the overall credit card operation, especially false applications, which are often difficult for card issuing banks to guard against . More and more college students are also joining the ranks of applying for credit cards. Sky-high tuition loans and daily living expenses have become the main reasons for college students to apply for credit cards . Credit card fraud is mainly divided into two categories: application fraud and transaction fraud. Application fraud mainly refers to the deceptive behavior of criminals using false identities, forging certificates or using other people's identities to apply for credit cards without consent . Transaction fraud is generally divided into counterfeit, mail-nonreceipt, lost or stolen, CNP fraud, and account takeover .

This study is to apply predictive model to detect credit card. Based on a dataset of credit card transaction, variable creation and feature selection are performed. Five machine learning algorithms are utilized to build the prediction model for credit card fraud detection. The model performances are analyzed and discussed through the experimental results .

2. Methodology

2.1 Data description

In this paper, a Tennessee government agencies explores a dataset of credit card transactions and build up a predictive model based on the dataset . There are 96,753 records in the notebookduring2010, with 1059 fraudulent activities. The file contains the following data, as presented in Table 1:

The study first processed the null entries, reducing the data volume entries, such as outliers. However, the remaining data volume is still huge, and it cannot be analyzed intuitively. It still needs visualization tools to make the data clearer. Figure 1 visually shows the top 15 common merchant descriptions, and Figure 2 depicts the top 10 of merchant in the states.

Table 1. Fields and Records

Data	Description
Record	Each data record's individual identifier. The time order is also represented by this field
Cardnum	The transaction's account number (the account numbers start with the digits 54, indicating that they are Mastercard transactions)
Date	The transaction's date. simply the month, day, and year; no hour of the day
Merchnum	a merchant identification number that is typically 12 digits long
Merch Description	A succinct text area for the merchant's description, usually 20 characters or less
Merch Zip	The retailer's zip code
Transtype	A code identifying the transaction type
Amount	The transaction's monetary value
Fraud	A description of the transaction that specifies whether or not it was fraudulent

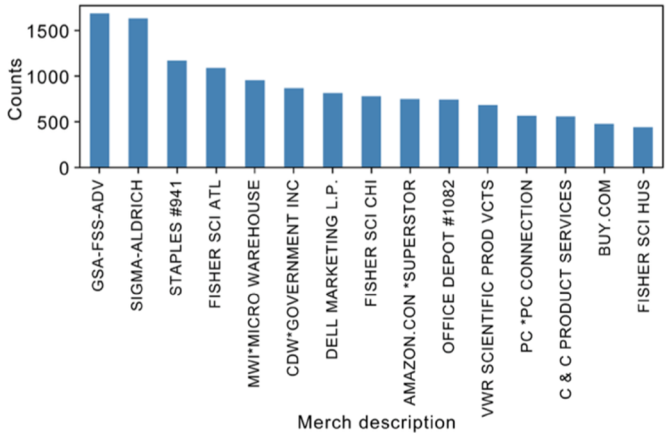

Fig. 1 Top 15 Common Merchant Descriptions

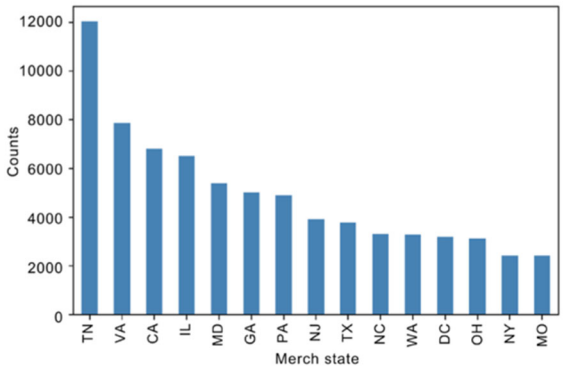

Fig. 2 Top 10 of the Most Frequently Observed Merchant States

This study will mainly focus on feature engineering first, which is also known as variable creation. Later, there will be some suggestions toward variable creation .

2.2 Variable creation

Variable creation is one of the most important processes to the entire setup of the predictive model because it wants to take all necessary factors into account and eliminate irrelevant ones. Variable creation, also known as feature engineering, uses raw data to demonstrate the different estimated

models for the given dataset . Hence, it will create four types of expert variables based on the future algorithm, as presented in Table 2:

Table 2. Variable Creation

Variables	Descriptions
Amount Variables	The specialty of these kind of variable is based on its timely indicator of unusual activity
FrequencyVariables	Frequency variables could directly illustrate how many transactions have been made throughout the day or during a period of time
Days SinceVariables	If there is a gap between two transaction date, it might use it as an indicator of fraudulent transaction.
Velocity Variables	This is similar to frequency variable, but it will use these variables to compare a large amount of transactions for a day and the average transaction during a specific time period.

2.3 Feature selection

Filter, wrapper, and embedded methods are the three major categorized methods which have been used in feature selection process. The study will explore the first two methods.

In the majority of cases, filter methods are considered as a pre-estimating step. Feature Selection is usually considered as a process of selecting significant factors that may affect the potential response for the model . Correlation among coefficients also plays an important role in the model. Based on the table, we could easily distinguish the different relationships among the various models. We use superviused and unsupervised machin learning model for respone variable and predictor variables. Here is the table illustration below:

Table 3. Table for defining correlation coefficients

Response Variable	Predictor Variables	
Feature\Response	Continuous	Categorical
Continuous	Pearson's Correlation	LDA
Categorical	Anova	Chi-Square

It attempts to use a subset of features and train a model utilizing them in wrapper methods. The study chooses to add or subtract features from the subset based on the conclusions that are derived from the prior model. Large computational calculation is also required during the process of the model establishment. There are numbers of algorithms involving exponential and integrals.

To determine the goodness of fit, Kolmogorov-Smirnov formula is utilized. It contrasts the variable's cumulative distribution function with a predetermined distribution. Kolmogorov-Smirnovtest is commonly used for binary classification problem. In short, it is called Kolmogorov-Smirnov test as KS method. Formula 1 and Formula 2 shows the mathematical description.

$$KS = max_x \int_{x_{min}}^{x} [P_{good} - P_{bad}]\, dx \tag{1}$$

$$KS = max_x \sum_{x_{min}}^{x} [P_{good} - P_{bad}]\, dx \tag{2}$$

2.4 Model Algorithm

2.4.1 Artificial neural network (ANN)

This technique uses the data that has been collected to forecast outcomes. The input layer, hidden layer, and output layer are the three layers that make up this system. Each layer is interconnected and

capable of communicating with one another to create effective results . There are different kinds of fraudulent activities, which require abundant storage for computational calculations .

$$S(x) = \frac{1}{(1+e^{-x})^2} \quad (3)$$

$$S'(x) = \frac{1}{(1+e^{-x})^2} = S(x)(1\text{-}S(x)) \quad (4)$$

In ANN, it will use this nonlinear model to construct the model . The study shows that all the linear combinations have been built through each layer and node . Eventually, it would construct the model through each layer and down to each node and combine all of them together to establish the model for a better decision.

2.4.2Support vector machine (SVM)

It is a supervised algorithm that has been used to predict the suspicious ongoing transactions, which may usually consider as fraud. Both classification and regression are carried out by it. This approach is not appropriate for huge data sets with significant noise . It could have a better vision for the SVM model .

2.4.3 Logistic regression

Logistic regression models use independent occurrence of findings to calculate the probability of an event occurring. This algorithm is employed to improve system performance. It combines the weak classifiers, as mentioned before, to achieve an efficient result . Based on the independent findings, the model could easily locate the fraudulent activities and mark them as special ones.

2.4.4 Random forest

Random Forest is a classification problem, which mainly composed of decision trees and voting process . The most commonly used one is called decision trees, which is an algorithm to filter the answer with the corresponding or designated conditions to fulfill the requirements .

2.4.5 Boosted tree

Extreme Gradient Boosting, also known as XGBoost tree, is a machine learning algorithm, distributed gradient-boosted decision trees (GBDT) .

3. Results

3.1 Model performance

Based on the prediction model results, there could have a clear picture of logistic regression model results. Logistic regression results, training dataset result, test dataset result all reflect through the charts below . For number of iterations, the result shows that 10.0000 for the final answer. Based on the logistic regression model, we could easily find out the corresponding values for each transformation. Take AIC, BIC, and Log-Likelihood as an example, each transformation brings a different result value before deciding the actual value.

Table 4. Logistic Regression Results

Description	Value	Description	Value
Model	Logit	Pseudo R-squared	0.538
Dependent variable	y	AIC	3382.435
Date	2019/2/1 19:57	BIC	3572.495
No. observations	62957	Log-Likelihood	−1670.2
Df Model	18	LL-Null	−3619.4
Df residuals	62957	LLR p-value	0.0000
Converged	1.0000	Scale	1.0000

3.2 Prediction result

Table 5 shows the training dataset results, testing dataset results, and OTT dataset results. Each figure clearly presents the records with the different features' results. Relevant data value also reflects in each column to specify the results . Training results and test results have been revealed through the model building process and automatic results . As mentioned before, it is easily to apply similar strategies for training dataset and test dataset.

Table 5. Training Results

Description	Value	Description	Value
Model	Logit	Pseudo R-squared	0.538
Dependent variable	y	AIC	3402.9056
Date	2019/2/1 19:57	BIC	3574.8652
No. observations	62957	Log-Likelihood	−1670.2
Df Model	18	LL-Null	−1682.5
Df residuals	62957	LLR p-value	0.0000
Converged	1.0000	Scale	1.0000

The testing dataset in Table 6 could also use assemble techniques to have different transformation for the model.

Table 6. Testing Results

Description	Value	Description	Value
Model	Logit	Pseudo R-squared	0.538
Dependent variable	y	AIC	3264.9856
Date	2019/2/1 19:57	BIC	3683.4756
No. observations	62957	Log-Likelihood	−1570.3
Df Model	18	LL-Null	−1721.6
Df residuals	62957	LLR p-value	0.0000
Converged	1.0000	Scale	1.0000

Based on Table 7, the study shows a clearly revealed table for all the records, which contains good and bad records in the dataset .

Table 7.OOT Results

#Records	#Goods	#Bads	Fraud Rate
12,427	12,248	179	0.0144

4. Conclusion

This paper explores non-traditional methods to distinguish a different way of building credit card fraud predictive models. The collected data is limited due to the lack of fraudulent labels. Also, the study shows that the boosted tree model is the best compared to the other 5 models mentioned in the paper. This study discusses several improvements based on variable selection process, which creates the new related concepts for it.

The study discusses all of the variable change and feature engineering process. This study discusses possible enhancement in some steps to improve our prediction model. It will focus on data collection, variable creation, and feature selection areas.

For Data collection section, researchers can make changes on data collection step to improve the model prediction by collecting some possible relevant data. The dataset used in this project contains 10 fields. Hence, it can improve credit card fraudulent pattern prediction accuracy by collecting more data fields. A few candidates are category of purchased items, the comparison of the order shipping address vs credit card billing address, and shipping method if it is an online order. Fraudsters have

some common favorite products, and they like to buy items that are easy to resell, or items with the best values, such as electronics, and fashion items.

Most fraudulent transactions have a different shipping address from the credit card billing address for online orders because the fraudsters need to ship the items to a different address other than card holders' billing address. Fraudsters usually prefer to use express shipping method, especially overnight. Regular consumers will consider overnight as very expensive unless it is necessary.

For Variable creation section, based on the new information, collected in data collection, itcould easily create some corresponding new variables. It could help a better understanding of the concept, if some variables are added, including purchase_cat, addresscomp (1 if addresses are the same, 0 if not), order_type (online or in-store), and shipping method.

For category purchased variable section, the variable contains the category information that buyers purchased. If the category information is not available, it can be collected by the product descriptions, and then the keywords in the product description are analyzed to get the category information.

For billing address and shipping address comparison flag variable section, it is a binary variable by comparing the credit card's billing address and the order's shipping address. The study should design 1 if the two addresses are the same, and 0 if the two addresses are inconsistent.

For order type variable section, it will set it to 1 if it is an online order and set it to 0 if it is an in-store purchase, which also means a credit card was present at the time of purchase.

For shipping method variable section, most merchants offer USPS, FedEx and UPS. Some merchants such as Amazon have their own shipping couriers. In general, all of them can be categorized as regular shipping, 2-day shipping, or overnight shipping. The study will show a value to each of these shipping methods in the next stage to better illustrate it.

For Feature Selection section, after the analysis of these variables, it will make a decision on what new feature variables to be included in the model.

With all of these improvements, the study would have a better model building process in variable selection and apply them in algorithm.

References

[1] Gao, J., Zhou, Z, Ai, J. Xia, B. and Coggeshall, S. (2019) Predicting Credit Card Transaction Fraud Using Machine Learning Algorithms. Journal of Intelligent Learning Systems and Applications, 11, 33-63. doi: 10.4236/jilsa.2019.113003.

[2] Shen, A., Tong, R. and Deng, Y., 2007, June. Application of classification models on credit card fraud detection. In 2007 International conference on service systems and service management (pp. 1-4). IEEE.

[3] Chaudhary, K., Yadav, J. and Mallick, B., 2012. A review of fraud detection techniques: Credit card. International Journal of Computer Applications, 45(1), pp.39-44.

[4] Srivastava, A., Kundu, A., Sural, S. and Majumdar, A., 2008. Credit card fraud detection using hidden Markov model. IEEE Transactions on dependable and secure computing, 5(1), pp.37-48.

[5] Bahnsen, A.C., Aouada, D., Stojanovic, A. and Ottersten, B., 2016. Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, pp.134-142.

[6] Seeja, K.R. and Zareapoor, M., 2014. Fraudminer: A novel credit card fraud detection model based on frequent itemset mining. The Scientific World Journal, 2014.

[7] Ogwueleka, F.N., 2011. Data mining application in credit card fraud detection system. Journal of Engineering Science and Technology, 6(3), pp.311-322.

[8] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C. and Bontempi, G., 2017. Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems, 29(8), pp.3784-3797.

[9] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S. and Jiang, C., 2018, March. Random forest for credit card fraud detection. In 2018 IEEE 15th international conference on networking, sensing and control (ICNSC) (pp. 1-6). IEEE.

[10] Raj, S.B.E. and Portia, A.A., 2011, March. Analysis on credit card fraud detection methods. In 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET) (pp. 152-156). IEEE.

[11] Awoyemi, J.O., Adetunmbi, A.O. and Oluwadare, S.A., 2017, October. Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 international conference on computing networking and informatics (ICCNI)(pp. 1-9). IEEE.

[12] Trivedi, N.K., Simaiya, S., Lilhore, U.K. and Sharma, S.K., 2020. An efficient credit card fraud detection model based on machine learning methods. International Journal of Advanced Science and Technology, 29(5), pp.3414-3424.

[13] Chen, T., 2014. Introduction to boosted trees. University of Washington Computer Science, 22(115), pp.14-40.

[14] Elith, J., Leathwick, J.R. and Hastie, T., 2008. A working guide to boosted regression trees. Journal of animal ecology, 77(4), pp.802-813.

[15] Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M., 2002. Logistic regression (p. 536). New York: Springer-Verlag.

[16] DeMaris, A., 1995. A tutorial in logistic regression. Journal of Marriage and the Family, pp.956-968.

[17] Sperandei, S., 2014. Understanding logistic regression analysis. Biochemia medica, 24(1), pp.12-18.

[18] Biau, G. and Scornet, E., 2016. A random forest guided tour. Test, 25(2), pp.197-227.

[19] Lin, W., Wu, Z., Lin, L., Wen, A. and Li, J., 2017. An ensemble random forest algorithm for insurance big data analysis. Ieee access, 5, pp.16568-16575.

[20] Schonlau, M. and Zou, R.Y., 2020. The random forest algorithm for statistical learning. The Stata Journal, 20(1), pp.3-29.

[21] Zheng, A. and Casari, A., 2018. Feature engineering for machine learning: principles and techniques for data scientists. " O'Reilly Media, Inc.".

[22] Li, Z., Ma, X. and Xin, H., 2017. Feature engineering of machine-learning chemisorption models for catalyst design. Catalysis today, 280, pp.232-238.

[23] Sirén, K., Millard, A., Petersen, B., Gilbert, M.T.P., Clokie, M.R. and Sicheritz-Pontén, T., 2021. Rapid discovery of novel prophages using biological feature engineering and machine learning. NAR genomics and bioinformatics, 3(1), p.lqaa109.

[24] Dai, D., Xu, T., Wei, X., Ding, G., Xu, Y., Zhang, J. and Zhang, H., 2020. Using machine learning and feature engineering to characterize limited material datasets of high-entropy alloys. Computational Materials Science, 175, p.109618.

[25] Xu, Y., Hong, K., Tsujii, J. and Chang, E.I.C., 2012. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. Journal of the American Medical Informatics Association, 19(5), pp.824-832.

[26] Chen, Z., Zhao, P., Li, F., Marquez-Lago, T.T., Leier, A., Revote, J., Zhu, Y., Powell, D.R., Akutsu, T., Webb, G.I. and Chou, K.C., 2020. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Briefings in bioinformatics, 21(3), pp.1047-1057.

[27] Roe, K.D., Jawa, V., Zhang, X., Chute, C.G., Epstein, J.A., Matelsky, J., Shpitser, I. and Taylor, C.O., 2020. Feature engineering with clinical expert knowledge: a case study assessment of machine learning model complexity and performance. PloS one, 15(4), p.e0231300.

[28] Lou, R., Lalevic, D., Chambers, C., Zafar, H.M. and Cook, T.S., 2020. Automated detection of radiology reports that require follow-up imaging using natural language processing feature engineering and machine learning classification. Journal of digital imaging, 33(1), pp.131-136.