Comparing ML Algorithms on Financial Fraud Detection

Chung Min Tae
FPT University
Hanoi, Vietnam
taemse0104@fpt.edu.vn

Phan Duy Hung
FPT University
Hanoi, Vietnam
hungpd2@fe.edu.vn

ABSTRACT

The problem of Financial Fraud has reached an alarming scale nowadays. Losses due to the fraud are reaching billions of dollars every year. To reduce it, decision systems that use efficient fraud detection algorithms should be invented. With the support of modern technologies, these systems are able to manage to analyze the information and to create a prediction feature model. However, the invention of these systems is not a trivial matter but a quite challenging task due to the huge amount of different and unbalanced data. Moreover, it is not clear which machine learning algorithm should be implemented. Therefore, our research is conducted to answer the question: which is the most suitable algorithm for the dataset in this research, especially when dealing with the large amount of uncleaned data.

CCS CONCEPTS

• Applied computing → Secure online transactions

KEYWORDS

Financial fraud, Machine Learning Algorithms, Balancing dataset, R language.

ACM Reference format:

Chung Min Tae and Phan Duy Hung. 2019. Comparing ML Algorithms on Financial Fraud Detection. In *Proceedings of 2019 2nd International Conference on Data Science and Information Technology (DSIT'19). Seoul, Republic of Korea, 5 pages.* https://doi.org/10.1145/3352411.3352416

1 Introduction

1.1 Problem and Motivation

Nowadays, credit card online attack became more and more popular. The PwC global economic crime survey in 2018 performed that 49% of organizations experienced economic crime, which dramatically raised from 36% in 2016 [1]. Those results reveal clearly that although the millions of dollars were being spent to tackle it, economic crime remains a persistent and serious issue which needed more researches to tackle. As stated in [2], in the last years, many studies have been performed using data mining to investigate new techniques to detect fraud on the basis of the fraudulent paths and different algorithms have been developed to block fraudulent transactions before they are filled. Nevertheless, main proportion of these studies just mentioned to a specific algorithm and comparison between two or three techniques while some other researches did not balance the dataset before apply algorithms. Therefore, in this paper, we aim to compare the performance of 7 supervised machine learning algorithms in detecting credit card fraud after balancing and preprocessing the available labeled dataset.

1.2 Related Work

In the last years, many studies have been performed using data mining to investigate new techniques to detect fraud on the basis of the fraudulent paths and different algorithms have been developed to block fraudulent transactions before they are filled. However, the large proportion of them did not conduct balancing the dataset before apply algorithms. In [3], decision trees and support vector machines are applied on a dataset obtained from a real world national bank's credit card data warehouses. They found out that decision trees outperform support vector machines in solving the problem. The authors in [4] developed two models based on logistic regression and support vector machines. The results show that model with filtered variables is the most suitable one to detect financial statement fraud one year before the outburst of fraud event and support vector machine with all variables is the most suitable one to detect financial statement fraud two year advance to the outburst of fraud event. Fraud detection models based on the decision trees was developed in [5] and founded that decision trees suffer from under fitting problem in case of imbalanced data set (case of fraud detection dataset). The research [6] investigates the performance of naive bayes, k-nearest neighbor and logistic regression on highly skewed credit card fraud data. The comparative results show that k-nearest neighbor performs better than naive bayes and logistic regression techniques. The paper [7] aimed to provide a comprehensive review of Hidden Markov Model (HMM) and Neural Networks (NN) techniques to detect credit card fraudulent in an effective way. This paper concluded that If one of these HMM or NN or combination of both the algorithm is applied in credit card fraud detection system, the probability of fraud transactions can be predicted and prevented from the unauthorized user access. In [8], three different machine learning algorithms namely logistic regression, decision tree, and self-organized map were used in order to train a model for the credit card fraud detection. The conclusion can be drawn that the decision tree model performs

DSIT '19, July 19–21, 2019, Seoul, Republic of Korea.
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7141-4/19/07…$15.00
https://doi.org/10.1145/3352411.3352416

better with unbalanced data without any pre-processing of that data on the fraudulent transactions classification problem.

As can be clearly seen, these examples mentioned above just referred to whether a specific machine learning technique or comparison between maximum three algorithms. Moreover, there is an example of research that did not balance and also preprocess the dataset before applying the algorithms. Thus, in this paper, we take into account 7 different supervised machine learning techniques and also include balancing and preprocessing the dataset to make the results more trustful.

1.3 Contributions

The main contribution of this paper is to provide the overall and reliable comparison of supervised machine learning techniques on credit card fraud detection by considering a quite large number of recent popular algorithms. In addition, this work provides the formalization of sampling methods adopted in unbalanced classification tasks. Based on ROC curve, we conclude that Synthetic Minority over-sampling Technique (SMOTE) gained the highest performance from data obtained as compared to other sampling methods. Moreover, this paper can be referred as a reference for data preprocessing which involves removing the fake and manage the transaction values.

The remainder of this paper is structured as follows: Chapter 2 aims to perform the project implementation including introduction about environment, dataset, dataset preprocessing, dataset balancing, machine learning algorithms performance, and finally comparison among these 7 algorithms. Meanwhile, Chapter 3 concludes which algorithm has higher performance result compared to other supervised machine learning algorithms.

2 Implementation

2.1 Environment

The machine that we use to perform research's algorithms has the following hardware properties:

- OS : Window 10 (x86-64)
- CPU : Intel CORE i5 2-CORE,
- RAM and Disc : 8GB and 250 GB

Besides, Software properties are also described as follows:

- IDE : Rstudio 1.1.463
- Tool : R version 3.5.1 (2018-07-02)

2.2 Dataset

The dataset is used from the public source https://www.kaggle.com/mlg-ulb/creditcardfraud.

Credit Card Fraud detection dataset contains transactions made by credit cards in September 2013 by European cardholders, which occurred in two days period. Additionally, this dataset is highly unbalanced, in detail, there are 284,807 rows (transactions) and only 492 are positive examples (frauds) which account for around 0.172% of all data. Features V1, V2 … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Time variable mentions to the seconds between each transaction and the first transaction in dataset. Amount variable means the transaction amount. Finally, Class is the data classification: the transaction is marked as (1) positive if it's fraud and marked as (0) negative if is genuine. Table 1 is a description of each feature.

Table 1: Feature in dataset

Feature	Description	Type
Time	Seconds Elapsed from first transaction	Numeric
V1 to V28	Anonymized information with PCA applied	Numeric
Amount	Transaction Amount	Numeric
Class	The actual classification classes (Boolean 0 or 1)	Numeric

2.3 Data Preprocessing

After having overall view, we perform an Exploratory Data Analysis (EDA) to gain detail understanding about this dataset. At first we check whether the relationship between Time variable and Class variable exists or not. After implementing some functions, we have the results that there is no relationship existing between Time variable and Class variable. In another word, we can cross out the Time variable, remaining the result of classification. Next step is identifying the association between principal components and class by preparing scatter plot matrix on training dataset. After that, by using the data but null we perform the density plot to understand whether the variable is normally distributed or not. Finally, we conduct correlation plot matrix to investigate the dependence between multiple variables from V1 to V28 at the same time, drawing a conclusion that every feature provided is essential and will not be reduced.

2.4 Data Balancing

There are many ways of dealing with imbalanced data, which distinguish into data and level algorithms. Among these techniques, the preferred one is data level method in which the analyst interacts with data as preprocessor to modify dataset, rebalance the unbalanced data and remove noise between two classes before using data in the algorithms. There are 5 main methods belong to data level technique that we will consider in this paper namely Over-sampling, Under-sampling, combined class method, SMOTE, and ROSE. Based on ROC (AUC), we summarize the result in the table 2 as follows:

Table 2: Table Captions

Items	AUC
Original dataset	0.786
Over-sampling	0.944
Under-sampling	0.775
Both-sampling	0.944
ROSE	0.928
SMOTE	**0.950**

Hence, we get the highest ROC curve from data obtained using SMOTE-sampling in comparison with other sampling methods. This technique combined with a more robust algorithm (random forest, boosting) can lead to exceptionally high accuracy. As the results, we will use SMOTE-sampling method to balance the dataset before applying 7 mentioned ML algorithms.

2.5 Machine Learning Algorithms

Firstly, we provide the overall view about procedure in Figure 1. Based on that, after generating the data through SMOTE-Sampling, in this section, we will build the model and apply it to the respective test data set. Then, we compute the Confusion Matrix, containing starting from up-left cell in clockwise the values: True Positive TP, false Positive FP, True Negative TN, and False Negative FN, and calculating Performance Parameters.

Accuracy and F1 score are key concepts in performance evaluation. Besides, Precision and Recall are tools served for calculation. When using Performance Measure with Accuracy, it is effective when the data is balanced. That is, it is advantageous to measure when the amount of data input for each class is the same. If the data is imbalanced, it may not be accurate when the input is different for each class. In other words, even though the performance of the classifier is not good, it can conclude that it is good. So, in the case of imbalanced data, when using Performance Measure with the F1-Score, the data is effective.

Since we created balanced data through SMOTE-Sampling, based on theory, we can just based on accuracy to evaluate the performance of 7 supervised algorithms. However, in this work, we made a performance measure based on both accuracy and also F1-score. The reason is that we want to double check between two evaluation methods and also provide a sufficient reference for other researchers. Below are the performances of 7 mentioned algorithms base on the Confusion matrix.

2.5.1 K-Nearest Neighbor Algorithm. K-Nearest Neighbor algorithm is a type of supervised learning that uses labeled data to perform classification [9]. It is an algorithm that is used for classifying problems during learning. Classification problem refers to the problem of classifying which group of existing data belongs to when new data comes in. When applying the K-NN algorithm, data must be standardized. After running the model, we gain this result that K-Nearest Neighbor result in Accuracy of 0.95945 and F Measure score of 0.97927.

2.5.2 Logistic Regression Algorithm. The second supervised learning technique that we mention in this research is Logistic Regression. Since logistic regression is used when the dependent variable is categorical data, the classification is classified into a specific category [10]. The dependent variable is a continuous measurement variable such as sales or profit. After performed R commands, we have summarized the results in the table above. We can see that Logistic Regression Algorithm result in Accuracy of 0.97037 and F Measure score of 0.98494.

2.5.3 Naive Bayes Algorithm. Then, we concern about Naive Bayes which is a simple technique for creating classifiers, trained using algorithms based on general principles rather than training through a single algorithm. In the probability model, the Naive Bayes classification is very efficient in Supervised Learning environment [11]. After running R commands, we can see that Naive Bayes result in Accuracy of 0.97091 and F Measure score of 0.98522.

2.5.4 Decision Tree Algorithm. Next, we care about technique called Decision Tree. Decision trees generate models represented by trees and rules. Decision trees are used for both classification (classification trees) and numeric prediction (regression trees) problems [12]. The two best-known and most widely used decision tree systems: (1) CART (Classification and Regression Trees) by Breiman et al. (1984) Breiman L, Friedman JH, Olshen RA, Stone, CJ (1984) Classification and Regression Trees (Wadsworth, Belmont, CA). (2) C4.5 by Quinlan (1993) Quinlan JR (1993) C4.5: Programs for Machine Learning (Morgan Kaufmann, San Mateo CA). Decision tree algorithms begin with the entire training dataset, split the data into two or more subsets according to some splitting criteria, and then repeatedly split each subset into smaller subsets until a stopping criterion is met. After performing R commands, we see that Decision Tree result in Accuracy of 0.95227 and F Measure score of 0.97551.

2.5.5 Random Forest Algorithm. After that, we turn to another supervised learning technique called Random Forest. The easiest to understand is random forests (RF) is to use multiple trees. Several trees form a forest. It can be expressed as a sequential partition of the data subsets and can be used for classification and regression prediction [13]. Decision trees are widely used in machine learning. The nature of the tree allows it to grow deeper and have higher discretion, but it has the problem of overfitting if it goes wrong. Random Forest is a model that creates multiple trees and averages these overfittings to reduce errors. After applying R commands, we see that Random Forest result in Accuracy of 0.98966 and F Measure score of 0.99479.

2.5.6 AdaBoost Algorithm. One more algorithm mentioned is Ada Boost. Boosting is a simple classifier to build a robust classifier. It is a way to create a strong classifier that combines weaker classifiers with slightly lower prediction performance and performs better. Adaboost is a combination of Adaptive and Boosting. This method is staggered (sequential) learning so that the weak classifiers complement each other and combine them to amplify the performance of the final strong classifier [14]. Gradient Boosting is Residual Fitting. After building a very simple model, we create a fitted model to Residual and combine the two. Then, when the residual model comes back from the combined model, the model that is fitted to the residual model is created, and it is repeatedly made to make the final model. After running R commands, we see that Ada Boost result in Accuracy of 0.96898 and F Measure score of 0.98422.

2.5.7 Neural Network Algorithm. Neural network algorithm methods include Feedforward, Backpropagation, Forward phase, and Backward phase. Feedforward is a method of calculating input layer, hidden layer, and output layer in order. Backpropagation is a typical algorithm for learning a feedforward neural network with multiple hidden layers with labeled learning data [15]. After performing R commands, we see that Neural Network result in Accuracy of 0.95186 and F Measure score of 0.97530.

2.6 Comparison

To conclude, we have the best evaluation of the accuracy of the Random Forest algorithm as 0.98400 and the F Measure score as 0.99194. Secondly, the AdaBoost algorithm has an excellent rating of 0.97090 and an F Measure score of 0.98522.

In addition, we make a comparison between our results with a similar source, which is performed in Figure 2. This reference also examines the performance of the same supervised machine learning algorithms in detecting credit card fraud, considering the same dataset. This research draws a conclusion that three algorithms, namely Random Forests, AdaBoost, and Neural networks, have higher performance than average predictive accuracy. Comparing with investigated reference, we have the same conclusion for Random Forests and AdaBoost algorithms performance. In terms of Neural Network, it has good performance in mentioned reference but not in our actual result because the parameters of this algorithm and the PCA should be considered more sufficiently.

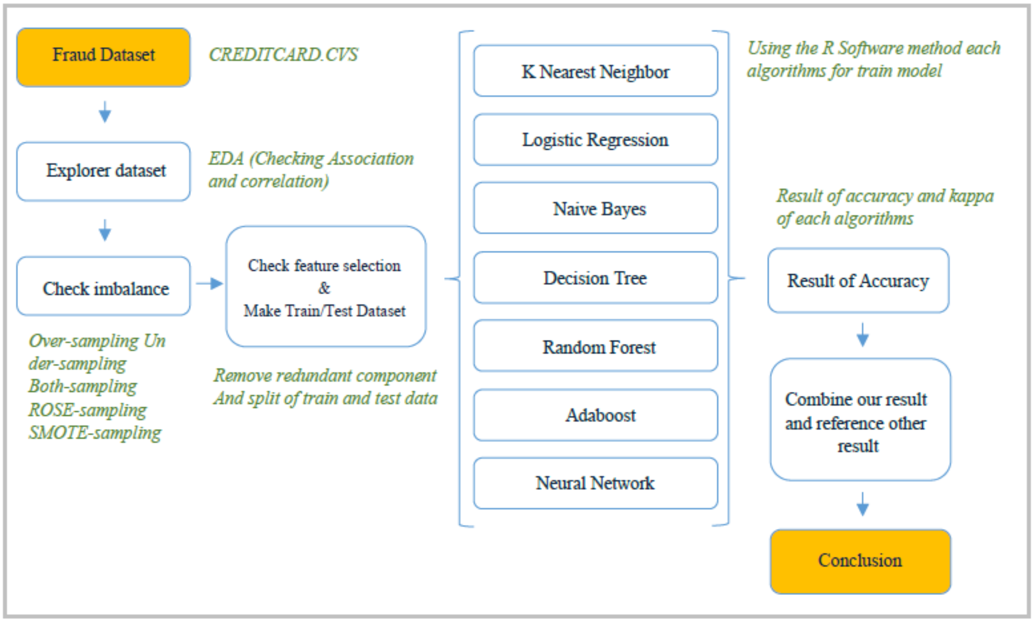
Fraud Dataset
CREDITCARD.CVS
Explorer dataset
EDA (Checking Association and correlation)
Check imbalance
Over-sampling Under-sampling
Both-sampling
ROSE-sampling
SMOTE-sampling
Check feature selection & Make Train/Test Dataset
Remove redundant component And split of train and test data
K Nearest Neighbor
Logistic Regression
Naive Bayes
Decision Tree
Random Forest
Adaboost
Neural Network
Using the R Software method each algorithms for train model
Result of accuracy and kappa of each algorithms
Result of Accuracy
Combine our result and reference other result
Conclusion

Figure 1: Work Flow for Comparison Fraud Detection Base on ML Algorithms

No	Algorithms Type	Our result of performance				Research result of performance (Source: https://www.dezyre.com)		
		Accuracy	Precision	Recall	F-Score	Average predictive accuracy	Training speed	Amount of parameter turning needed
1	Random forest	0.98400	0.99970	0.98430	0.99194	Higher	Slow	Some
2	Adaboost	0.97090	0.99975	0.97111	0.98522	Higher	Slow	Some
3	Logistic Regression	0.97040	0.99987	0.97044	0.98494	Lower	Fast	None
4	K-Nearest Neighbour	0.96900	0.99980	0.96912	0.98422	Lower	Fast	Minimal
5	Naive Bayes	0.95940	0.99987	0.95950	0.97927	Lower	Fast	Some for feature extraction
6	Decision Tree	0.95230	0.99978	0.95239	0.97551	Lower	Fast	Some
7	Neural Network	0.95190	0.99983	0.95194	0.97530	Higher	Slow	Lots

Figure 2: Result of Comparing Algorithm on Financial Fraud Detection and Predictive Accuracy.

3 Conclusion and Perspectives

To sum up, based on Accuracy and F1 score, we have the same conclusion that Random Forest algorithm is ranked as the best technique to detect credit fraud of mentioned dataset with Accuracy of 0.98400 and F1 score of 0.99194. The Random Forest algorithm is followed by AdaBoost algorithm which gained Accuracy of 0.97090 and F1 score of 0.98522.

In addition, we compared this result with other reference described in Figure 2. The result in that reference shows that there are three supervised machine learning techniques gaining better performance in comparison with other considered algorithms, namely Random Forests, AdaBoost, and Neural Networks. As can be clearly seen, our work and investigated reference all agree about the performance evaluation of Random Forests, AdaBoost algorithms.

In this work, the dataset that we take into consideration is already labeled; however, obtaining a labeled dataset in real situation is very difficult. For that reason, we will consider unsupervised techniques in the future research so that our result can be more easily applied in the real world. This paper also can give a reference to many field in Data Analytics, for example, Bioinformatics [16-18], Pattern Recognition [19-21], etc.

REFERENCE

[1] "PwC's Global Economic Crime and Fraud Survey 2018," 2018. [Online]. Available: https://www.pwc.com/gx/en/services/advisory/forensics/economic-crime-survey.html. [Accessed 7 November 2018].

[2] Lusis. [Online]. Available: https://www.lusispayments.com/uploads/4/4/8/2/44826195/a_comparison_of_machine_learning_techniques_for_credit_card_fraud_detection.pdf. [Accessed 7 November 2018].

[3] Sahin, Y., Duman, E. Detecting Credit Card Fraud by Decision Trees and Support Vector Machines, in Proceedings of the International MultiConference of Engineers and Computer Scientists 2011, 16-18 March, 2011, Hong Kong.

[4] Huang, S. Y. Fraud Detection Model by Using Support Vector Machine Techniques, Chiayi, Taiwan: International Journal of Digital Content Technology & its Applications, 2013.

[5] Ehramikar, S. The Enhancemeat of Credit Card Fraud Detectioa Systems using Machine Leaming Methodology, Master of Applied Science Thesis, University of Toronto, 2000.

[6] John, O. A., Adebayo, O. A., Samuel, A. O. Credit card fraud detection using machine learning techniques: A comparative analysis, ICCNI 2017: International Conference on Computing, Networking and Informatics, Lagos, Nigeria.

[7] Rajamani, R., Rathika, M. Credit Card Fraud Detection using Hidden Morkov Model and Neural Networks, in Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications, 2015.

[8] Anohhin, I. Data mining and machine learning for fraud detection, Master thesis, Tallinn, 2017.

[9] Shichao, Z. Efficient kNN Classification With Different Numbers of Nearest Neighbors. IEEE Transactions on Neural Networks and Learning Systems, Volume: 29, Issue: 5 , May 2018.

[10] Yue, Z. 2016. A Logistic Regression Based Approach for Software Test Management. In the Proceedings of the 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). IEEE, Chengdu, China.

[11] Mykhailo, G.,Volodymyr, M. 2017. Fake news detection using naive Bayes classifier. In the Proceedings of the 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). IEEE, Kiev, Ukraine.

[12] Zhong, Y. 2016. The analysis of cases based on decision tree. In the Proceedings of the 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE, Beijing, China.

[13] Jitendra, K. J., Rita, S. 2017. Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. In the Proceedings of the 2017 World Congress on Computing and Communication Technologies (WCCCT). IEEE, Tiruchirappalli, India.

[14] Zhe, L., Xin, D., Yixin, C. 2017. Label confidence based AdaBoost algorithm. In the Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, Anchorage, AK, USA.

[15] Bogdan, M. W. 2017. Neural network architectures and learning algorithms. In the Proceedings of the 2009 IEEE Industrial Electronics Magazine.

[16] Hung, P. D, Hanh, T. D., and Diep, V. T. 2018. Breast Cancer Prediction using Spark MLlib and ML packages. ICBRA, 5th International Conference on Bioinformatics Research and Applications, 12, 2018, Hong Kong.

[17] Hung, P. D. 2018. Detection of Central Sleep Apnea based on a single-lead ECG. ICBRA, 5th International Conference on Bioinformatics Research and Applications, 12, 2018, Hong Kong.

[18] Hung , P. D. 2018. Central Sleep Apnea Detection Using an Accelerometer. In Proceedings of the 2018 International Conference on Control and Computer Vision (ICCCV '18). ACM, New York, NY, USA, 106-111.

[19] Nam, N. T., Hung, P. D. 2018. Pest detection on Traps using Deep Convolutional Neural Networks. In Proceedings of the 2018 International Conference on Control and Computer Vision (ICCCV '18). ACM, New York, NY, USA, 33-38.

[20] Hung, P. D., Linh, D. Q. 2019. Implementing an Android Application for Automatic Vietnamese Business Card Recognition. Pattern Recognition and Image Analysis, ISSN 1054-6618 29 (1), 203-213.

[21] Phan, P. D., Giang, T. M., Nam, L. H., et al. 2019. Vietnamese speech command recognition using Recurrent Neural Networks. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019.