Multiple Machine Learning Models on Credit Card Fraud Detection

Minjun Dai*

Department of Statistics&Economics, University of Toronto, Toronto, Canada

*Corresponding author: minjun.dai@mail.utoronto.ca

Abstract. Nowadays, there is a huge increase in digital financial fraud as a result of the widespread usage of credit cards for online purchases. Therefore, a credit card fraud detection method with high accuracy, minimizing the risk of losing money when the transaction occurs, becomes imperative. The study first processed a synthetic data set, discarding useless features, using one hot encoding to convey categorical information to numerical ones and separating the training and testing datasets. Based on the new formed datasets, the study built three Machine Learning (ML) models for fraud detection, which are the Support Vector Machine (SVM) model, the logistic regression model and the decision tree model, respectively. Next, the study implemented on training these three models by fitting the model on training datasets and predicting the test data in testing datasets. Lastly, the heatmap named confusion matrix was provided to visualize the outcomes, indicating the accuracy of each model. Also, the study uses another four performance measurement scores which are Area under the Receiver Operating Characteristic Curve (ROC AUC score), accuracy classification score, balanced F1 score and precision score for evaluating models. Comparing these three models based on the confusion matrix and the four accuracy scores, the study explored that the decision tree model can achieve the best performance compared to other models in predicting fraud.

Keywords: Credit card fraud detection; machine learning; SVM; logistic regression; decision tree.

1. Introduction

Fraud refers to the crime of obtaining goods, services and money by deceiving people. Due to the wide use of credit cards for online purchases, there's an explosion in digital financial fraud nowadays. Credit card fraud, simply, happens when the credit card of an individual is used by another individual without his awareness that the card is being used. Using phishing emails, phony websites, bogus mobile apps, fake social media profiles, and other mechanisms to obtain information and defraud consumers and businesses illegally are all criminal methods in digital fraud. Therefore, a reliable fraud detection becomes imperative which can detect scams, and thus minimize the risk of losing money when the fraudulent transaction occurs. Based on existing purchasing data, credit card fraud detection is a process that creates a fraud risk score along with real-time monitoring of fraudulent events by adaptive and predictive analytics. Credit card fraud detection is actually difficult based on the problem that lacking public available datasets on financial services due to privacy reasons. Moreover, the main challenge of fraud detection is the imbalance of the data set, as the fraudulent transactions accounts for a very tiny percentage of all transactions.

In recent years, various machine learning-based systems have been implemented to identify different types of fraud due to the success of artificial intelligence in other domains [1-4], but there exists limitation on each detection model. Arjun Joshua suggests in his Kaggle notebook that the first technique considered for dealing with imbalance data before applying a machine learning (ML) algorithm is to under sample the frauds which is reducing the number of instances or over sample the non-frauds by copying the number of instances of non-frauds, for example using synthetic minority oversampling technique SMOTE [5]. Then the neural network was implemented on the processed data set after over sampling and under sampling.

However, Janio Martinez Bachmann states in his Kaggle notebook the neural network on the oversampled dataset predicts fewer correct fraud transactions than our model using the undersample dataset [6], while under sampling is unable to detect for a large number of cases non fraud transactions correctly and instead, misclassifies those non fraud transactions as fraud cases. If this happens, some

normal cards of real customers are blocked, where these customers will complain and not rely on the certain bank or corporation any more, causing a large capital loss of the financial markets. Next this study takes outlier detection into consideration and carry out the outlier removal on our oversample dataset, improving the accuracy of the detection. Previously undiscovered types of fraud may be detected under outlier detection because it's unsupervised methods. However, it still needs work to figure out how to deal with the outlier and ensure the accuracy of outlier detection. There are other methods considering ML models directly on imbalanced data that did not discuss and compare in detail. Arjun Joshua chooses ensembles of decision trees, because he discovered that such algorithms not only allow for constructing a model that can cope with the missing values, but they naturally allow for speedup in parallel-processing. But the deficiency is it has a degree of bias and is slightly underfitting.

To solve the limitation mentioned above, this study suggests three fraud detection models, which are the Support Vector Machines (SVM) model [7], the logistic regression model and the decision tree model [8, 9], and compares them for detail based on the performance of the model, discussing using which model in certain circumstance and how to train the model fitting the data.

2. Method

2.1 Dataset preparation

The study uses a synthetic data set, which is generated using the financial simulator called PaySim [10], simulating mobile money transactions based on an original dataset. This dataset consists of 6,362,620 data items and 11 feature columns which are step, type, amount, nameOrig, oldBalanceOrig, newBalanceOrig, nameDest, oldBalanceDest, newBalanceDest, isFraud, isFlaggedFraud. Step maps a unit of time in the real world. In this case, 1 step is 1 hour of time. And the total number of steps is 744 for this 30-days simulation dataset. And the predicted objective of the study is the type called "isFraud", where the fraud transactions are tagged 0 or 1 to show whether it is a fraud or not. There exists a highly imbalanced distribution between fraud and non-fraud transactions with just 0.002964 percent fraudulent transactions with respect to all transactions.

In order to deal with data and remain meaningful ones which are related to fraud transactions, the study explores each column of data and discards "nameOrig", "nameDest" entries which is the customer who starts the transactions and the recipients' ID of the transactions, and "isFlaggedFraud" entry, set from the "isFraud" where the amount of "Transfer" is over 200, 000, because of the insignificance. There remain 8 columns including step, type, amount, "oldBalanceOrig", "newBalanceOrig", "oldBalanceDest", "newBalanceDest", which is the balance of customers' accounts and recipients' accounts before and after the transactions, and "isFraud". Then, the study uses one hot encoding for conveying categorical information into numerical format, which can be easily fed into machine learning algorithms for processing. Next, in order to test the performance of various models accurately, the study randomly splits 80% of original datasets into training datasets and the 20% left over datasets into testing datasets.

2.2 SVM model

The first model this study presents is the Support Vector Machines (SVM) model, using linear support vector classification in scikit-learn to build the model with regularization parameter 1×10^9, creating a line or a hyperplane which separates the data into classes. Then the study trains the SVM model by fitting the model on the training datasets and predicting the test data on the testing datasets. In addition, the results were also visualized based on the heatmap named confusion matrix and the model's accuracy was evaluated by computing area under the Receiver Operating Characteristic Curve (ROC AUC score), accuracy classification score, balanced F1 score and precision score.

2.3 Logistic regression

The second model presented in this study is the logistic regression model, modeling the probability of an event taking place by using logistic regression classifier in scikit-learn. Then the study fits the model on the training datasets and predicts the test data in the testing datasets. The outcome is visualized by heatmap and named confusion matrix in the study. The model's accuracy is tested as well by ROC AUC score, accuracy classification score, balanced F1 score and precision score.

2.4 Decision tree

The third model this study suggests is the decision tree model, which creates a model that predicts the value of a target variable through learning decision rules inferred from the data features by using a decision tree classifier in in scikit-learn. In the decision tree model, the study choose entropy as the function to measure the quality of a split and used the "best" supported strategies choosing the best split and set the random state to be 32, controlling the randomness of the estimator. Next, for training the decision trees model, the model fitting of the training data set and the prediction of the test data set are implemented. Moreover, visualizing the result by heatmap and testing the model's accuracy by measuring ROC AUC score, accuracy classification score, F1 score and precision score is done in the study.

3. Result and discussion

After building three models for fraud detection, the study presents the confusion matrix by heatmaps and computes four accuracy criteria scores for each model. Fig. 1, Fig. 2 and Fig. 3 shown below are the confusion matrix for the SVM model, the logistic regression model and the decision trees model, respectively. Table 1 shows the value of each accuracy criterion for three models.

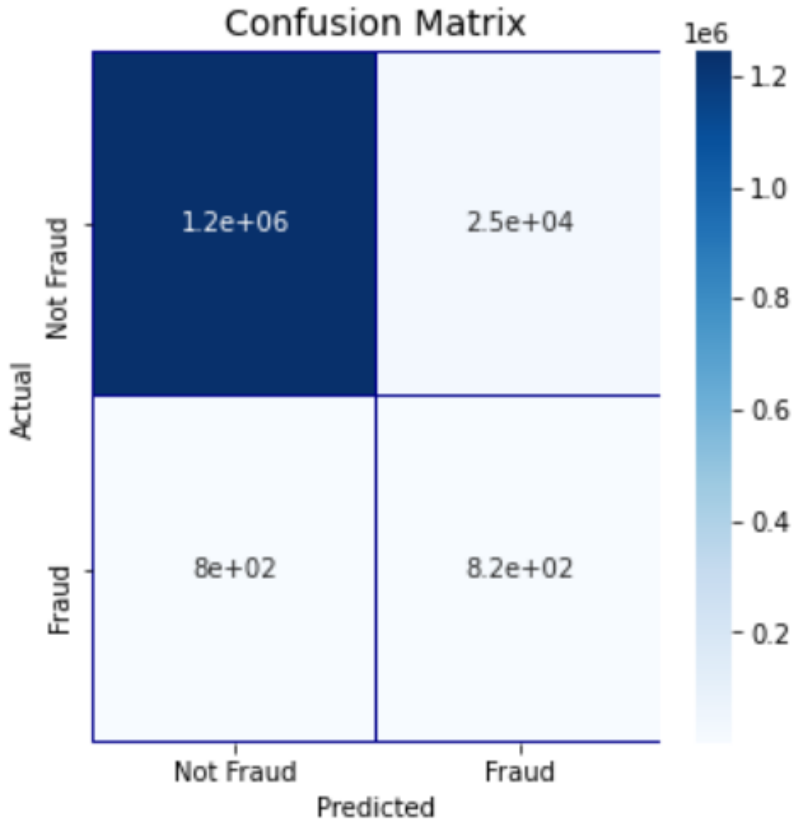

Fig. 1 The confusion matrix of the SVM model.

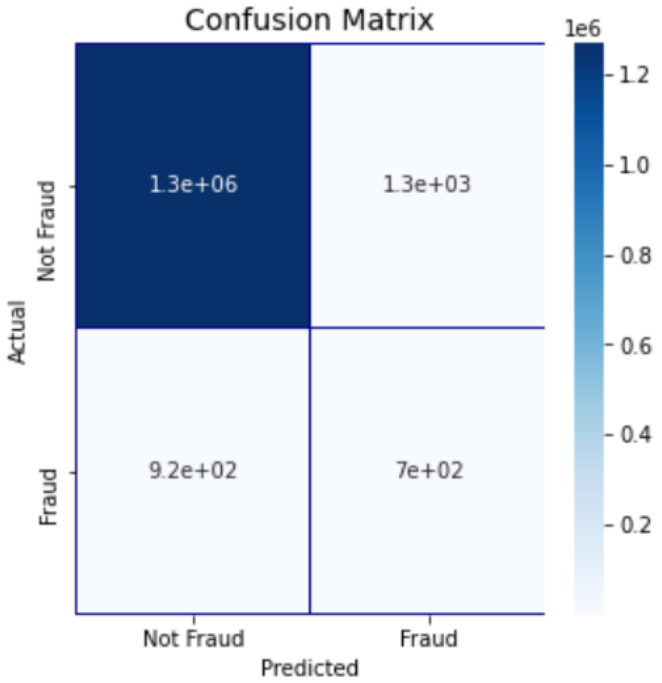

Fig. 2 The confusion matrix of the logistic regression model.

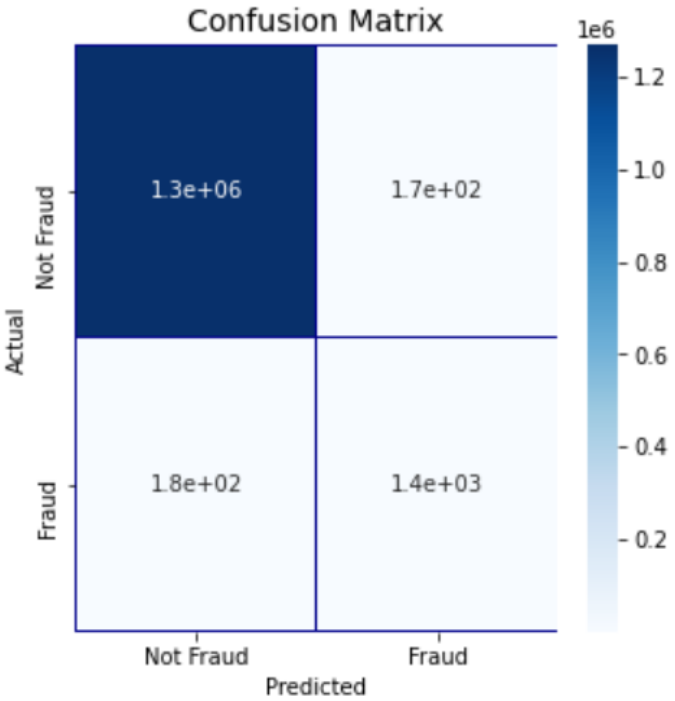

Fig. 3 The confusion matrix of the decision tree model.

Table 1. The performance of employed machine learning models.

Models\Criteria	ROC AUC score	Accuracy classification score	F1 score	precision score
SVM model	0.743	0.979	0.059	0.031
Logistic regression model	0.715	0.998	0.388	0.352
Decision tree model	0.944	0.999	0.891	0.895

Through these three figures, it can be observed that the results located in the diagonal line are correct predict: not fraud transactions are predicted to be not fraud transactions, fraud transactions are

predicted to be fraud transactions. On the contrary, the other two regions are not correct, which are not expected to be happened. Therefore, the study found that the classifier from the SVM model, logistic regression model to decision trees model becomes more accurate due to the numbers of correctly sorted transactions are increasing while the numbers of incorrectly sorted transactions are decreasing. Moreover, the accuracy scores the study computes shows the accuracy is enhanced as well, resulting from the value of each criterion is increasing. Especially, the value of the four accuracy scores for decision tress is very close to 1.

The result is within expectation. For the SVM model based on the linear kernel, it is so simple that cannot make a precise classification on a large, complicated datasets like the margin of separation between classes is not so clear. However, logistic regression model fit fraud detection better because the objective of the study is to detect whether the transaction is fraud or not (0 or 1). Although the logistic regression model is improved compared to the first one, it still not so accurate. The study decides to choose decision trees as the final model due to its highest accuracy. Unlike the first two models, decision tree model can group the transaction into more categories, which is more applicable for this complex dataset.

4. Conclusion

The study conducted credit card fraud detection by using three ML models directly on imbalanced data, discussed these three models and compared them in details based on the performance of the model. In this work, the SVM model, the logistic regression model and the decision tree model were built for predicting fraud. Among these three models, decision tree model performed the best that the study chose at last for fraud detection due to the highest accuracy scores it has, which means this model better predicts whether a transaction is fraud or not. However, this study just used and compared a few models. More well-established models e.g. neural networks on fraud detection can be added for further discussion.

References

[1] Kolachalama, Vijaya B., and Priya S. Garg. Machine learning and medical education. NPJ digital medicine 1.1, 2018, 1-3.

[2] Zhong, Z, et al. Class-specific object proposals re-ranking for object detection in automatic driving. Neurocomputing 242, 2017, 187-194.

[3] Yu, Q, et al. Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. China Communications, 17.3, 2020, 46-57.

[4] Bakr, Mohammed Abdul Hay Abu, et al. Breast cancer prediction using JNN. International Journal of Academic Information Systems Research (IJAISR) 4.10, 2020.

[5] Kaggle. Predicting Fraud in Financial Payment Services. 2018. https://www.kaggle.com/code/arjunjoshua/predicting-fraud-in-financial-payment-services/notebook

[6] Kaggle. Credit Fraud || Dealing with imbalanced Datasets. 2019. https://www.kaggle.com/code/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets/notebook

[7] Noble, William S. What is a support vector machine? Nature biotechnology 24.12, 2006, 1565-1567.

[8] LaValley, Michael P. Logistic regression. Circulation 117.18, 2008, 2395-2399.

[9] Myles, Anthony J., et al. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society 18.6, 2004, 275-285.

[10] Lopez-Rojas, Edgar, Ahmad Elmir, and Stefan Axelsson. PaySim: A financial mobile money simulator for fraud detection. 28th European Modeling and Simulation Symposium, EMSS, Larnaca. Dime University of Genoa, 2016.