Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230934

Comparison of machine learning algorithms for credit card fraud transaction prediction

Jindi Wu

School of Computer Science and Technology, East China Normal University, Shanghai, 200062, China

10205102490@stu.ecnu.edu.cn

Abstract. Nowadays, credit card transaction is widely used in commercial activities, the number of fraud transaction is increasing. Especially when e-commerce and online shopping have arisen explosively and COVID-19 pandemic spread, the fraud transaction increasing surprisingly. Detecting the credit card fraud transaction is of great importance for both banks and depositors. To verify the effectiveness of machine learning algorithms for this task, several models are leveraged and examined in this work. These models include naive Gaussian bayes model, logistic regression model, random forest classifier, convolutional neural network (CNN) and support vector machine (SVM). Besides, the effectiveness of pre-processing techniques, including undersampling and oversampling, is also conducted on the training dataset for validating their effectiveness on predicting whether a transaction is a fraud or not and their joint effort with different models. In this work, these models are implemented with European dataset of credit card fraud and the best method is selected for this application.

Keywords: machine learning, fraud transaction, random forest, convolutional neural network.

1. Introduction

Credit card is widely seen in nowadays commercial systems. It generally refers to a credit card that allows them to buy services and goods within credit limit [1]. Transaction fraud is encountered by nearly all business entities, especially these companies with online payments [2, 3]. Fraudster could make purchases online using stolen credit card numbers, which will harm the right of both card holders and companies. Terribly, with the rapid development of economy, the number of the fraud transactions is increasing surprisingly. According to a financial website named technote global, since the COVID-19 pandemic exploded, the fraud costs increased 10%-16% across APAC from 2019 pre-pandemic levels [4, 5]. From the Figure 1, it can be seen that before the pre-pandemic period, the fraud transaction has already grown quickly.

DOI: 10.54254/2755-2721/6/20230934

Figure 1. U.S fraud detection.

So, it is necessary to look for a proper way to predict the transaction fraud [6, 7]. Or even predict the likelihood of someone is will be involved in a transaction fraud event. Several machine learning (ML) methods are chosen to train the dataset and make some predictions. As can be seen, it is a two-category problem. ML is an appropriate way to solve the classification problem like that. Multiple ML techniques are used like Naive Bayes Classification [8], Logistic Regression [9], Random Forest [10], SVM [11] and CNN [12] to train the dataset. And because the dataset is highly unbalanced, the universal ways undersampling and oversampling are used to implement unbalanced dataset and compare their training effect.

2. Introduction

2.1. Dataset

The dataset is come from Kaggle. In accordance with the introduction of the website, the dataset contains nearly 300,000 credit card transactions records in European. The record time is mostly from September 2013. Totally, there are 492 frauds out of 284,807 transactions. As can be identified, the dataset is highly skewed, where the frauds occupies about 0.172% of all transactions. The jupyter notebook is used to draw the composition of the dataset as shown in Figure 2.

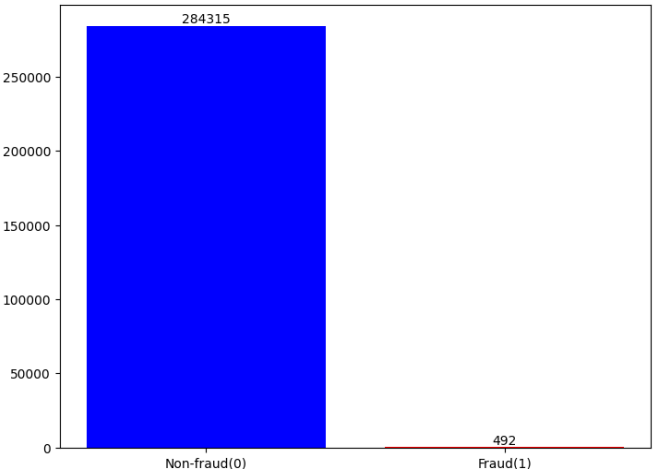

Figure 2. The bar graph of the dataset.

DOI: 10.54254/2755-2721/6/20230934

For protecting user indentities and sensitive features, some part of the dataset is reduced by PCA dimensionality reduction method. Anonymous features include 28 dimensions which are the principal components of PCA.

In this work, many classical models, such as the decision tree model, Gaussian naive bayes, random forest model, logistic regression model, and CNN models are used to train the dataset. Otherwise, to solve the problem that the dataset is unbalanced, undersampling and oversampling methods are applied to process the dataset to prompt the models' effect.

2.2. Pre-processing

2.2.1. Undersampling. Undersampling is frequently used method to reduce the imbalance of the dataset, where, for each training epoch, only part of the samples in the dominant class will be selected for training and all the samples in the minority class will be leveraged.

2.2.2. Oversampling. Oversampling is a more frequent method used to diminish the side effect caused by the imbalance of the dataset. Especially in the era of big data, abandon samples is a waste of previous data. This method will resample the images from the minority class for learning to make sure the numbers of samples in various classes are nearly the same. Dataset balanced by the undersampling and oversampling is demonstrated in Figure 3 respectively. As could be observed, both methods could generate a balanced dataset for learning.

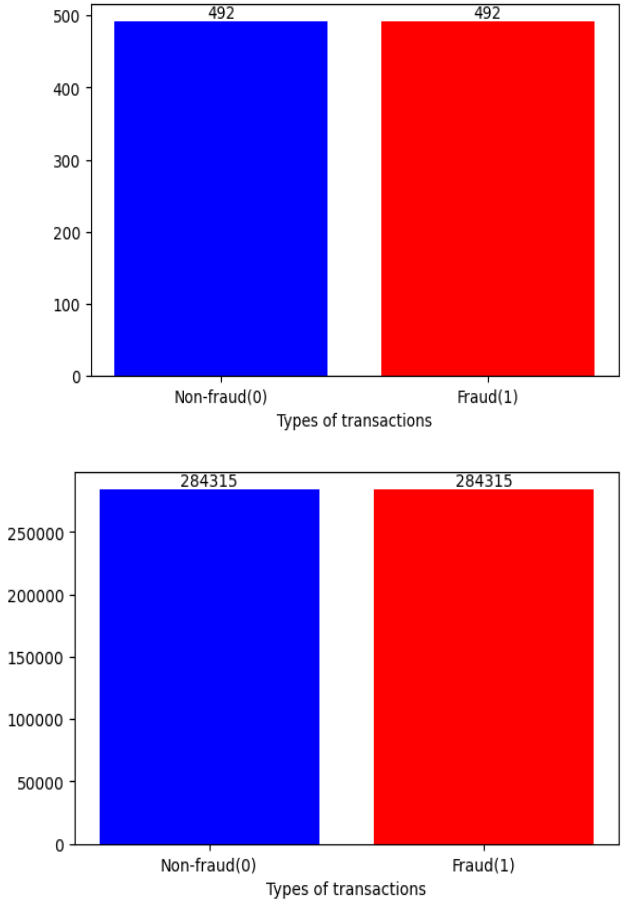

Figure 3. Dataset distribution after undersampling (left) and oversampling (right).

DOI: 10.54254/2755-2721/6/20230934

2.3. Models

2.3.1. Gussian Naive Bayes model. Gussian Naive Bayesian is one of the most classical generative probabilistic models for learning. It leverages the Bayesian law for learning. It assumes that all the features are independent to each other. In practical, this assumption largely decreases the computational consumption during training and inference. In this work, it is chosen as a representative probabilistic model for comparing the superiority of it among other discriminative classifiers.

2.3.2. Logistic regression model. Logistic regression is one of the earliest models for classification, which is widely used in binary classification tasks. It could output prediction in the form of the probabilities. This model, however, faces difficulties on solving the nonlinear classification problem, which is regarded as a weak model for machine learning.

2.3.3. Random forest classifier. Rnadom forest is composed a series of tree models for classification. It is a kind of boosting model, where the performance of the trees will be integrated as the result. It is a stable and powerful model. For a single tree model, the performance is sensitive to randomness, which is likely to be affected and hence generated unstable results. The random forest, however, could mitigate this drawback by aggregating different trees. Moreover, different from logistic regression, this model is good at handling discrete features as good as the continuous ones and hence suitable for learning from dataset with the combination of both continuous and discrete features.

2.3.4. SVM model. The objective of the SVM is to learn representative hyperplanes for classification. It is one of the most powerful classifiers before the deep era. The learned hyperplane could separate data points of different categories apart for classification. It is implemented by maximizing the distance between the data points of various classes. After training, only the representative support vectors are preserved for inference, which will dramatically decrease the computational cost, especially in the ear of big data. By using the kernel function, it could be adopted to the nonlinear problems.

2.3.5. CNN. CNN is widely used in today's machine learning applications. It achieves stat-of-the-art performance this field, such as the face recognition, art generation and so on. It is good at extracting complex features hidden beneath the dataset, by its hierarchical structures. The entire architecture is mainly composed by different operations, such as the convolutional operations, pooling operations, and activation operations. By using these operations, the input will go through them layer by layer and forms the final output features for classification. In multiple classification problem, the output is usually activated by a Softmax layer, to transfer the features to the probabilities. It is suitable for this problem.

3. Result

The result part demonstrates the performance of the models, by measuring the accuracy of different models and corresponding settings, such as undersampling and over sampling.

3.1. Performance

Models' accuracy of different models is shown in Table 1. Moreover, corresponding confusion matrix outputs of various models are demonstrated in Figure 4, 5, 6, 7 and 8, respectively.

Table 1. Accuracy performance of these models.

	Gaussian Naive Bayes	Logistic Regression Model	Random Forest Classifier	SVM Model	CNN network
Baseline Acc	97.84%	99.91%	99.96%	99.94%	99.95%
Undersampling Acc	97.66%	99.96%	99.96%	99.94%	99.82%
Oversampling Acc	97.57%	97.72%	99.96%	98.66%	99.90%

DOI: 10.54254/2755-2721/6/20230934

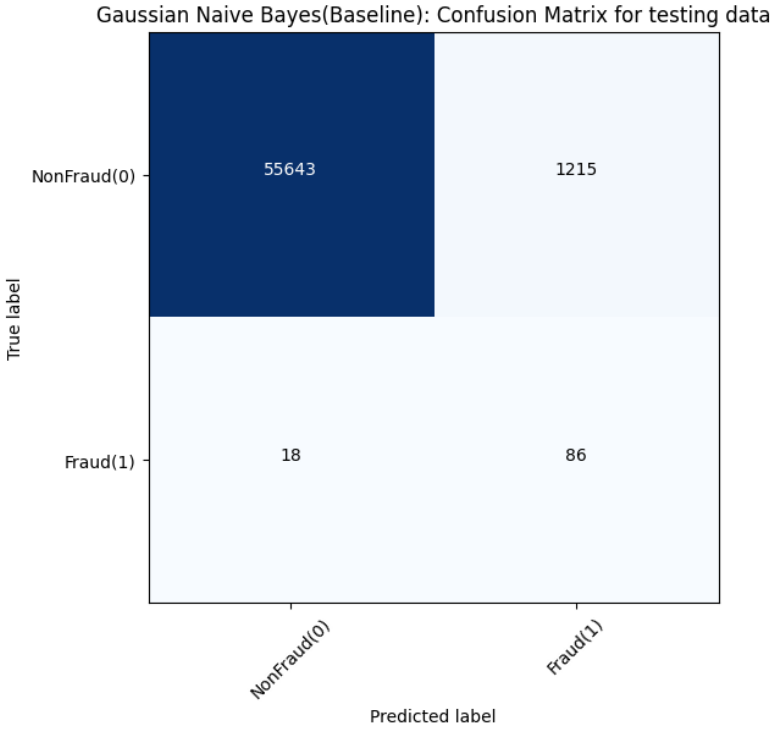

Figure 4. Gaussian naïve bayes result measured by confusion matrix.

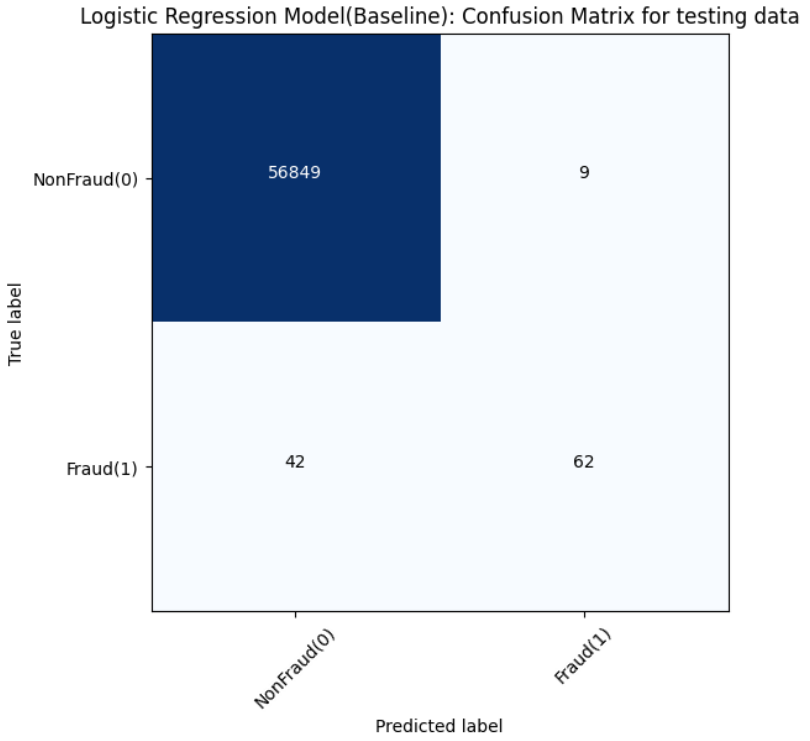

Figure 5. Logistic regression result measured by confusion matrix.

DOI: 10.54254/2755-2721/6/20230934

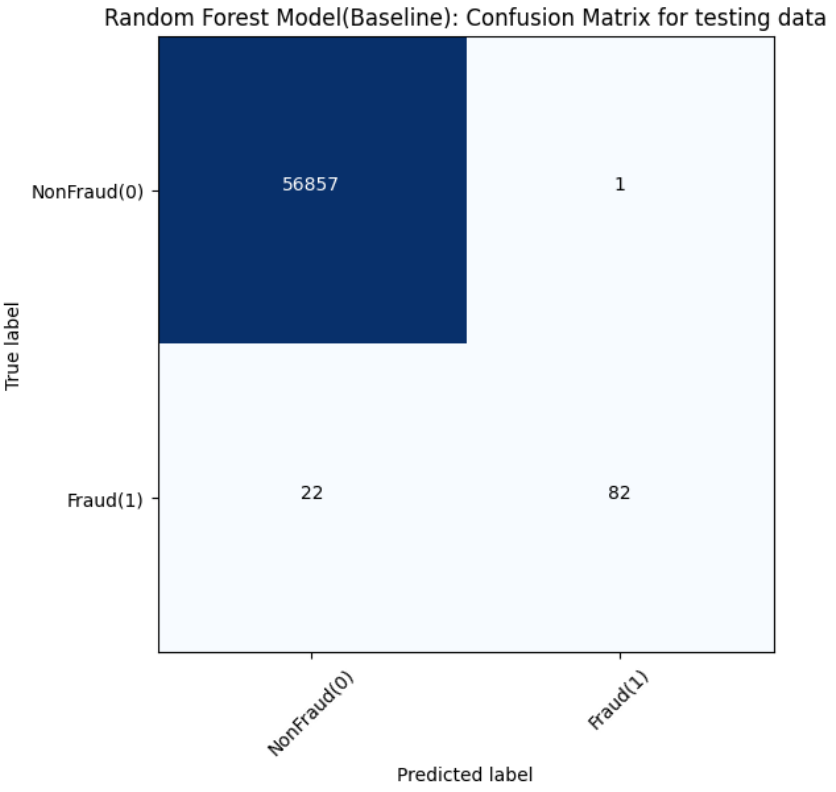

Figure 6. Random forest result measured by confusion matrix.

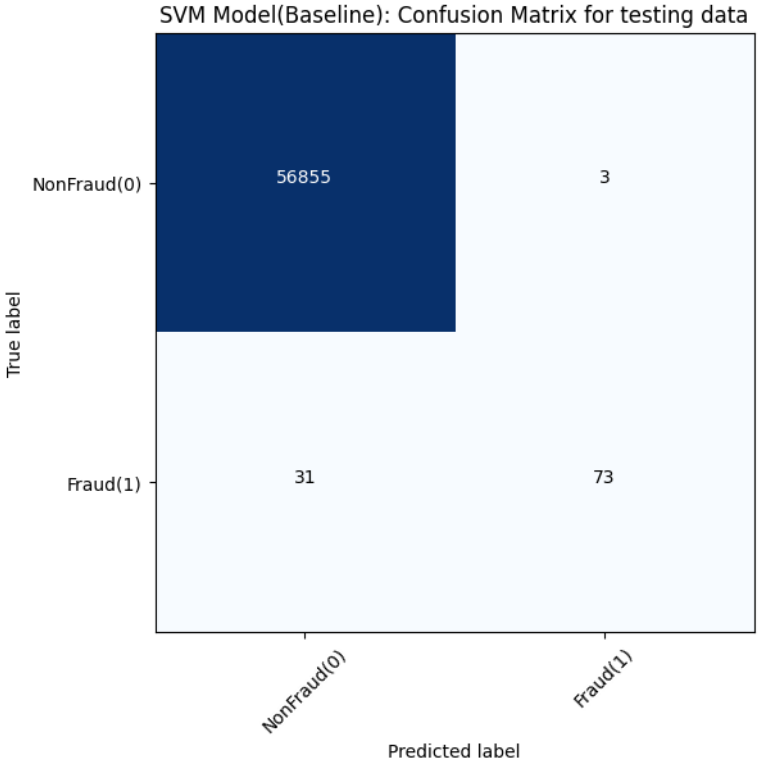

Figure 7. SVM result measured by confusion matrix.

DOI: 10.54254/2755-2721/6/20230934

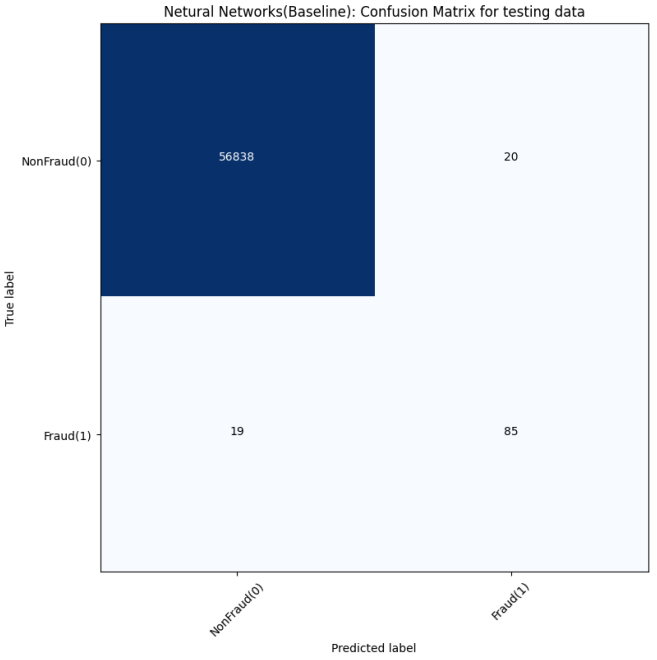

Figure 8. Neural network result measured by confusion matrix.

4. Discussion

In the baseline experiments, the Naive Gaussian Bayes Model can be trained in a relatively shorter time. However, it performs a lower accuracy than other models, only 97.84%. And other 4 models, they have similar effect. Other models all perform a nearly 100% accuracy.

In the undersampling experiments, the Naive Gaussian Bayes Model can be trained in a relatively shorter time. However, it performs a lower accuracy than other models, only 97.66%. And other 4 models, they have similar effect. Other models all perform a nearly 100% accuracy. The results of the undersampling settings are shown in Figure 9, 10, 11 and 12 respectively.

In the oversampling experiments, the most models perform unsatisfactory than other pre-processing situation: most of them perform a worse accuracy, especially CNN Model performs only 93.50% accuracy on test dataset. And in this part, Naive Gaussian Bayes Model can also be trained in a relatively shorter time. Different results related to oversampling settings are demonstrated in Figure 13, 14, 15, and 16 respectively.

Especially, when the CNN model is leveraged to train the dataset, the accuracy and loss curve is plotted by epoch elapsing. It could be found that the model would go convergence and keep stable after long epoch. So, maxpooling is used to increase the performance of the model. After processing, the inflection point decline from 20 to 15 epoches (undersampling) and 10 to less than 5 epoches. The situation is as following.

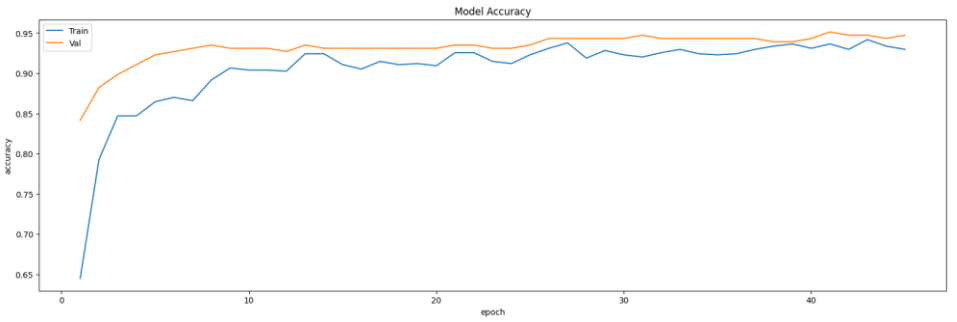

Figure 9. Curves of undersampling setting showing the model accuracy.

DOI: 10.54254/2755-2721/6/20230934

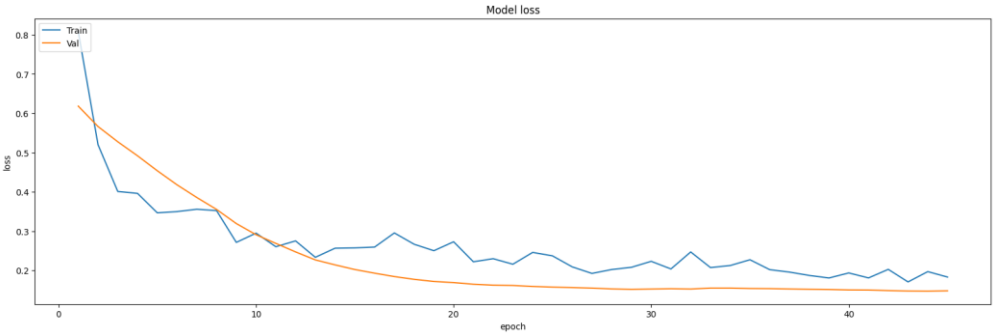

Figure 10. Curves of undersampling setting showing the model model loss.

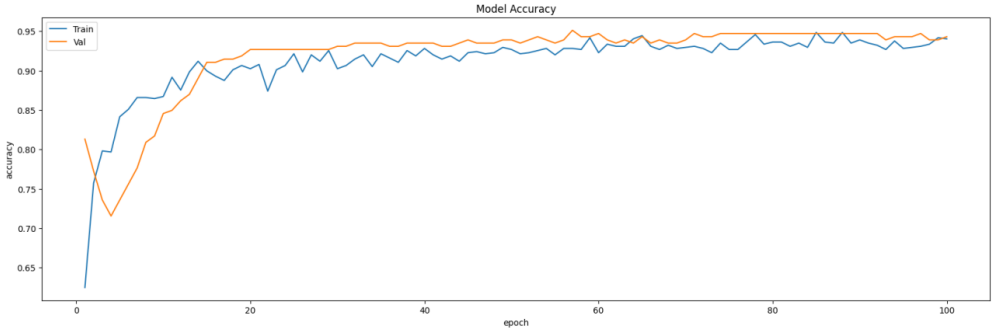

Figure 11. Curves of undersampling and max pooling setting showing the model accuracy.

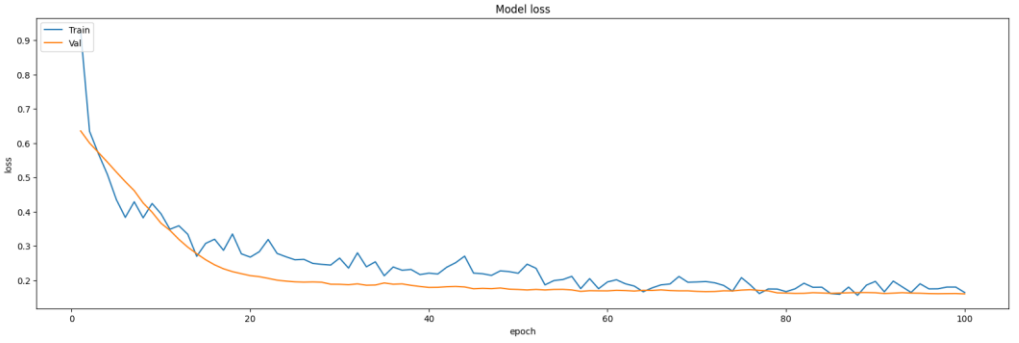

Figure 12. Curves of undersampling and max pooling setting showing the model loss.

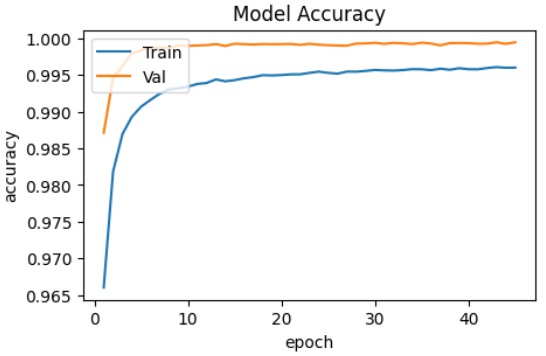

Figure 13. Accuracy of oversampling setting.

DOI: 10.54254/2755-2721/6/20230934

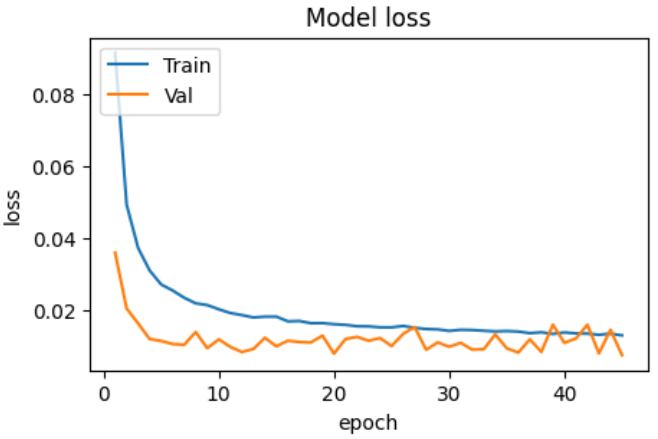

Figure 14. Loss of oversampling setting.

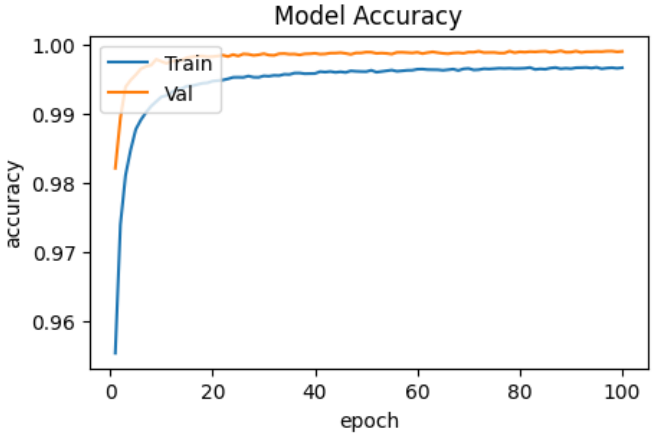

Figure 15. Accuracy of oversampling and maxpooling setting.

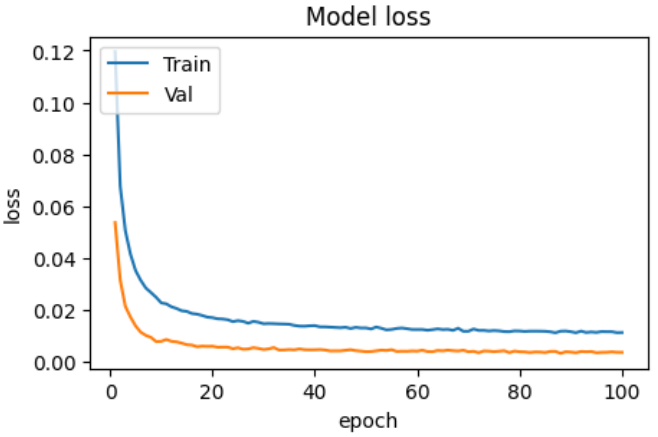

Figure 16. Loss of oversampling and maxpooling setting.

Furthermore, some other machine learning models could be implemented, especially netural network models to improve this work.

DOI: 10.54254/2755-2721/6/20230934

5. Conclusion

This article attempts to the problem to predict a credit card transaction is a fraud or not independently. And different models are used respectively, like Naive Guassian Bayes Model, Logistic Regression Model, Random Forest Classifier and CNN to train the dataset and test on test dataset. Because of the sufficient of the data, baseline and undersampling have no obvious difference. However, when the oversampling stategy is applied, some models' accuracy come to be a little fluctuated. Especially, this work emphasizes the part of CNN building and try to use maxpooling to improve its performance and received a satisfied conclusion. It shows that when jointly use the maxpooling and the oversampling, the training becomes stable, which is embodied as the smooth loss curve and the accuracy curve. This work through light on the comparison of different performances of the models, as well as the effectiveness of the undersampling and oversampling techniques. In the future, it could be used for the recognition of the transaction fraud. To further improve the performance, more advanced neural network architecture could be used for this task.

References

[1] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision support systems, 50(3), 602-613.

[2] Delamaire, L., Abdou, H., & Pointon, J. (2009). Credit card fraud and detection techniques: a review. Banks and Bank systems, 4(2), 57-68.

[3] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision support systems, 50(3), 602-613.

[4] Bandyopadhyay, S. K., & Dutta, S. (2020). Detection of fraud transactions using recurrent neural network during COVID-19: fraud transaction during COVID-19. Journal of Advanced Research in Medical Science & Technology, 7(3), 16-21.

[5] Alfaiz, N. S., & Fati, S. M. (2022). Enhanced Credit Card Fraud Detection Model Using Machine Learning. Electronics, 11(4), 662.

[6] Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. Procedia computer science, 165, 631-641.

[7] Adewumi, A. O., & Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. International Journal of System Assurance Engineering and Management, 8(2), 937-953.

[8] Ren, J., Lee, S. D., Chen, X., Kao, B., Cheng, R., & Cheung, D. (2009, December). Naive bayes classification of uncertain data. In 2009 Ninth IEEE international conference on data mining, 944-949.

[9] LaValley, M. P. (2008). Logistic regression. Circulation, 117(18), 2395-2399.

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

[11] Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.

[12] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.