Analysis of Fraud Detection Prediction Using Synthetic Minority Over-Sampling Technique

Uma Maheswari Ramisetty[1], Venkata Nagesh Kumar Gundavarapu[2(✉)], Akanksha Mishra[3], and Sravana Kumar Bali[4]

[1] Vignan's Institute of Information Technology, Visakhapatnam, India
[2] JNTUA College of Engineering Pulivendula, Pulivendula, India
drgvnk14@gmail.com
[3] Vignan's Institute of Engineering for Women, Visakhapatnam, India
[4] GITAM University, Visakhapatnam, India

Abstract. Credit cards are increasingly being used in real life for a wide variety of purposes. Because of the growing number of users, the number of scammers is also growing at an accelerating rate. E-commerce fraud detection methods are critical for reducing losses. Models developed in the past using unbalanced datasets show a high degree of accuracy. The precision, recall, and weighted average precision and recall are all quite low for the models. As a result of this research, techniques such as logistic regression (LR) and random forest (RF), along with SMOTE, were developed to increase the model's performance with imbalanced datasets. SMOTE techniques are used to balance the datasets because they are so unbalanced. SMOTE analysis has revealed that the RF with SMOTE is the best model for detecting credit card fraud, with accuracy, precision, and recall scores of 99.95%, 85.40%, 86.02%, and 85.71%, respectively.

Keywords: Synthetic Minority Over-sampling Technique (SMOTE) · eXtreme Gradient boosting · Accuracy · Precision

1 Introduction

The use of credit cards as a payment method has grown in recent years due to the convenience it provides. Credit cards are the most frequently used payment method for internet purchases. As the number of people who use credit cards increases, so does the number of those who fall victim to fraud. Today's businesses face a serious threat from credit card fraud. The use of a credit card without the consent or knowledge of the cardholder is referred to as credit card fraud. The cardholder and the card issuer are unaware that the card is being used. Card fraud occurs when one individual uses another person's credit card for personal reasons. Performing these illegal transactions may be done in order to avoid paying for items or to obtain an unauthorised account balance. These are only a couple of possibilities. When credit cards are stolen without the owner's knowledge or authorization, fraudsters profit personally. There is no price for using a virtual credit card, but you are need to supply additional information such as your

V. E. Balas et al. (Eds.): AITEES 2022, AHIS 3, pp. 3–12, 2023.
https://doi.org/10.2991/978-94-6463-074-9_2

credit card number, expiration date, and security code in order to use PayPal. In order to conduct a fake purchase, thieves only need access to the card transaction's details. Data from stolen credit cards, lost cards, forgeries, grabbing card data, and interception are just a few of the ways credit card information can be obtained. Fraudsters will always try to rationalise their actions, making it nearly impossible to discover fraudulent activity. Research and investigation into fraudulent transactions are required in order to stop them in the future. To detect this form of fraud, look for differences from the usual pattern on each card while calculating the costs. An abnormality in expenditure is studied further if it differs from the norm. Models for detecting credit card fraud are created by utilizing supervised machine learning techniques.

Several new algorithms have been created to identify literary forgeries by various writers in literature. DT, support vector machine, and RF classifier methods were utilised to construct the model suggested by Manohar S et al. [1]. The percentages of accuracy achieved were 99.7%, 99.8%, and 99.7%, respectively. Shirgave et al. [2] used a supervised machine learning RF technique to handle the class imbalance and idea drift problems and presented a ranking of alert methods based on probability to rank the alerts. Using the rating alert approach, fraud warnings can be identified and the fraudster's location tracked. An oversampling strategy was used by Lakshmi S. V. S. et al. [3] to balance the dataset. R and LR are used to create the model. The model was created by combining RF and DT classifiers, and the resulting accuracy levels are 90%, 94%, and 95%. As a result, the RF is regarded as the top model [4]. The model proposed by Bhanusri and colleagues was 100% accurate and used a radio frequency algorithm with a boost approach. The notion drift problem was overcome by Vaishnavi Nath et al. [5] via a feedback system. A one-class SVM is used to deal with the unbalanced dataset. The Local Outlier Factor and the Isolation Forest algorithm were demonstrated by S. P. Maniraj et al. [6] to predict credit card fraud. The model's accuracy is 99.6%, but its precision is just 28%, which is extremely poor due to the dataset's imbalance. Megasari et al. [7] developed a credit card fraud detection model with a 99.87% accuracy rate using an imbalanced dataset and the Isolation Forest technique. Nevertheless, no sampling approach to balance the dataset was taken into account. Navanshu et al. [8] led a team of researchers who built classifiers with 97.7% precision, 95.5% recall, and 98.6% efficiency to address the issue. When it came to accuracy, they discovered that RF had the best recall and efficiency. Masoumeh et al. [9] separated the dataset into four categories and used the data from each group separately to build a model for the imbalanced dataset. This paper's DT algorithm was implemented using the Nave Bayes classifier, the K-Nearest Neighbour algorithm, SVM, and the Bagging ensemble classifier. In order to discover the best answer, multiple researchers [10–12] worked on various optimization strategies. They haven't, however, taken machine learning into account. Using a Radial Basis Function kernel, the team led by V. Dheepa [13] came up with an SVM that had an accuracy of 80% and an error rate of 20%. Research done by Rashmi S and other people found the radio frequency (RF) algorithm to be 97% accurate at figuring out who isn't who they say they are. However, given the dataset, neither method is considered a sampling approach. It was found that Nave Bayes was the fastest, while LR was the most accurate. More et al. [14] proposed a system to classify fraud detection alerts

using Random Forest supervised learning technique to classify alert as fraudulent or non fraudulent.

While a lot of researchers have looked at different linear regression techniques for detecting credit card fraud, none have looked into the use of LR and RF with SMOTE. To improve accuracy, credit card apps might use LR and RF models. SMOTE has been modified to include the LR and RF to improve the model's precision. Utilizing RF, a model to prevent credit card fraud is constructed using the LR algorithm and SMOTE. The unbalanced dataset is resolved using the SMOTE. This study presents the results of the SMOTE-based predictive models.

In this paper, an effective SMOTE technique, coupled with RF and LR, was created to detect credit card fraud transactions when the dataset is substantially imbalanced and the models built using this imbalanced dataset are biassed towards the dataset's majority class distribution. Effective sampling procedures should be employed to balance the dataset because of the large number of credit card transactions in it. This will lead to better model performance. Because of under sampling, some significant information will be omitted from the model's training. Oversampling against undersampling has the advantage of generating new information while preserving data from the minority group and allowing the majority group an equal opportunity to show information. There should be technology to take on this problem in order to solve these difficulties. There are a lot of different ways to use SMOTE, like LR and RF, to get what you want to happen.

2 Proposed Methodology

LR and RF are supervised classification algorithms used to predict the target variable based on the values of the feature variables given to the model. Based on the number of categories in the target class, LR is divided into 3 types: binary or binomial, multinomial, and ordinal. Since our target class has 2 categories, i.e., 0 or 1, we are using binomial LR and RF. LR uses a sigmoid function which plots the classification probability in an "S" shaped curve within a range of 0 and 1. The default threshold value is set at 0.5. The values predicted above 0.5 are predicted as 'YES/1', and below 0.5 are predicted as 'NO/0'.

Figure 1 shows the block diagram for building a model using LR and RF by importing the required libraries to read the dataset. By importing several classifiers, the model is trained with test data and validation data, and then the new model is predicted with labelled data with new features. The accuracy of the LR model is evaluated by using an evaluation matrix, and the model's performance is measured. The procedure for building the model is given step-by-step.

From Fig. 2, which includes the software component SMOTE, the flowchart for the credit card fraud detection system architecture can be traced. The flowchart depicts the graphical depiction of the system architecture. The architecture was created with SMOTE in mind to detect credit card fraud. The training dataset is used to create a machine learning model. The model is trained using the training dataset, which has been preprocessed and separated into training and test sets. We must correct for the imbalanced dataset by drawing samples from it using the SMOTE approach. SMOTE balances the dataset by creating new synthetic instances of the minority class in the dataset while

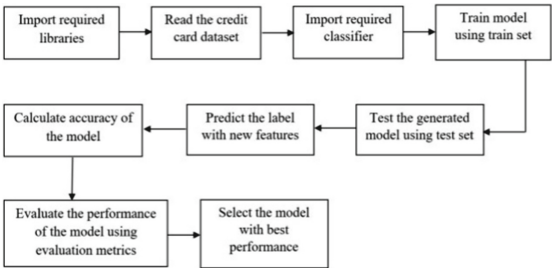

Fig. 1. Block diagram to build the model

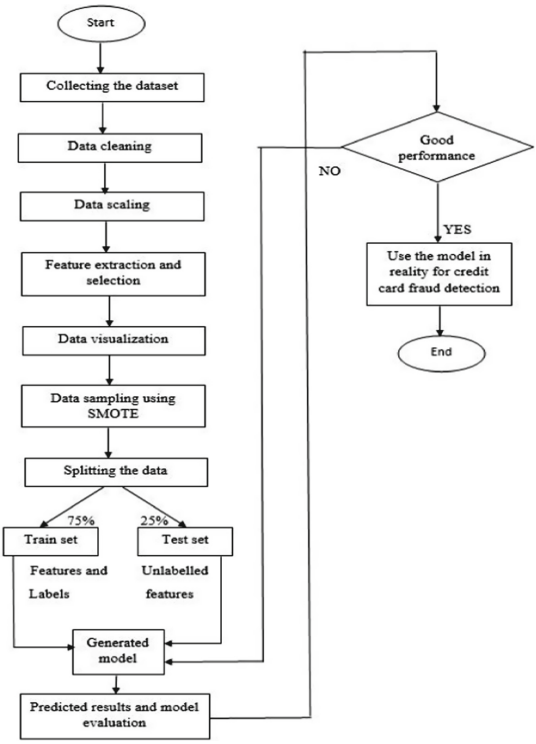

Fig. 2. Flow chart of the credit card fraud detection system with SMOTE

remaining true to the class-specific biases and distribution. The predictive model is built with LR, and the model's performance is measured using several assessment criteria. The general approach to building a model with SMOTE is the same whether SMOTE is used or not; the primary difference is that the dataset is sampled after it has been divided into the training and test sets.

3 Results and Discussion

These models are for the 284,807 transactions in the model, of which 0.172%, or 492, are fraudulent transactions, with the remaining 2,269,379 being non-fraudulent. In other words, the dataset is severely skewed. The model is trained for 75% of the dataset and tested for the remaining 25%. Because the dataset is highly skewed, the model will classify transactions incorrectly, resulting in the model's predicted accuracy not being met. As a consequence, we balanced the dataset using SMOTE. It improves the model's performance. Confusion matrix, accuracy, precision, recall, and F1 score are among the evaluation measures used to assess the model's performance. The models were created using LR and RF, and their performance was evaluated.

Figure 3 represents the time feature distribution over legitimate transactions. The time feature distribution over legitimate transactions has a cyclical distribution, which means that there is no unusual behaviour of users during legitimate transactions. Figure 4 shows the time feature distribution over fraud transactions, which has many time variations among all the fraud transactions. The distribution of time features over fraud transactions is more even, which shows how fraudsters act during fraud transactions.

Figure 5 shows the amount of feature distribution over legitimate transactions, which is highly right-skewed, meaning it has a peak value at the beginning but the rest of the graph becomes flat. It indicates that there are a greater number of outliers in the number of features. Figure 6 shows the amount of feature distribution over fraud transactions, which

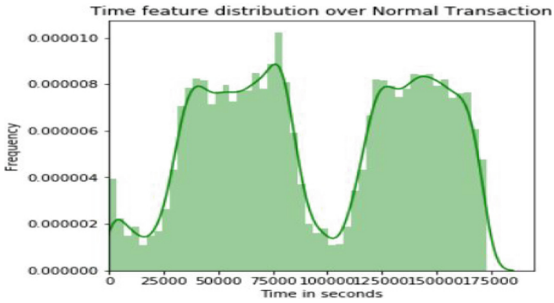

Fig. 3. Time distribution over legitimate transaction

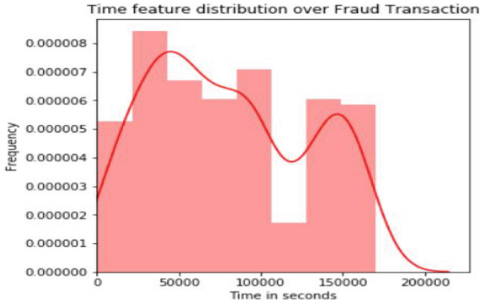

Fig. 4. Time distribution over fraud transaction

is slightly skewed with a few outliers. It shows that there are more unusual transactions than fraud transactions, which is how you can tell them apart.

Figure 7 exhibits the class distribution of the credit card transaction dataset, which is balanced with 199008 legitimate transactions and 199008 fraud transactions after

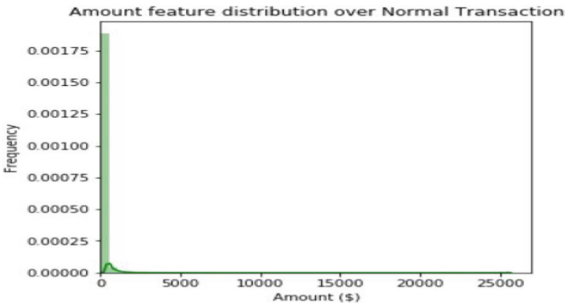

Fig. 5. Amount distribution over legitimate transaction

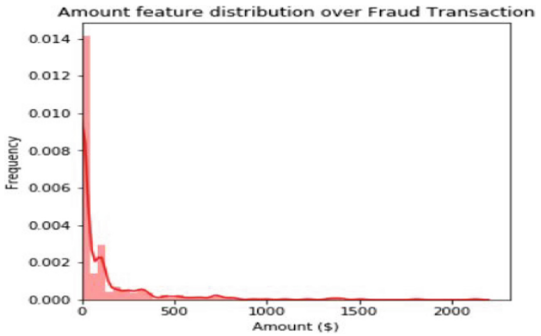

Fig. 6. Amount distributions over fraud transaction

Fig. 7. Class distribution with SMOTE

applying SMOTE. Because of applying the SMOTE method to the credit card transaction dataset, which has a total of 284,807 transactions, of which 492 of the transactions are fraudulent transactions and the remaining are non-fraud transactions, it has been converted into a balanced dataset with 199008 non-fraud transactions and 199008 fraud transactions.

Figure 8 represents the confusion matrix of LR which predicted 83021 TN, 2286 FP, 10 FN, and 126 TP transactions. Figure 9 represents the confusion matrix of the RF which has predicted 85287 TN, 20 FP, 19 FN, and 117 TP transactions. By comparing Fig. 8 and 9, we can say that in the LR model, 2286 transactions were misclassified as fraud even though they were non-fraud transactions, but in the RF model, only 20 transactions were misclassified as fraud even though they were non-fraud transactions. Misclassifying the transactions reduces the performance of the model.

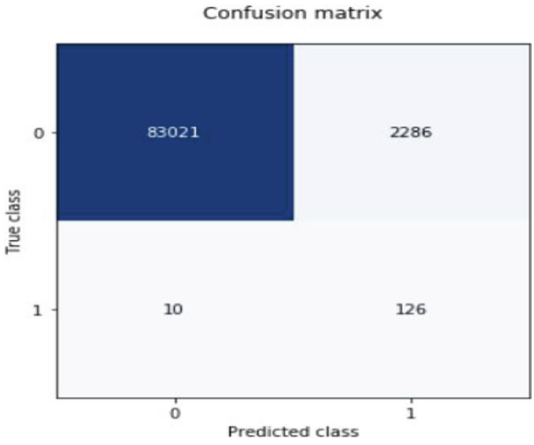

Fig. 8. Confusion matrix of LR

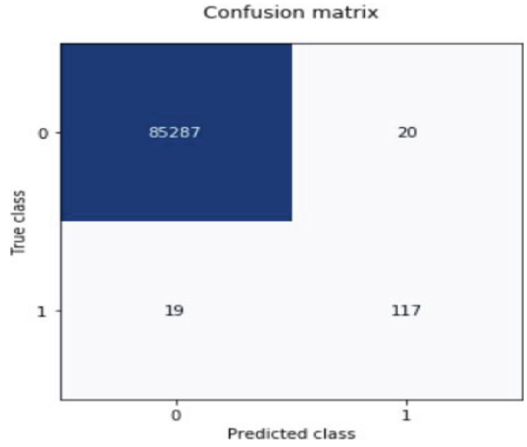

Fig. 9. Confusion matrix of RF

LR had a ROC curve shown in Fig. 10. On the horizontal axis is represented the True Positive Rate (which is represented as the horizontal length of the y-axis) and on the vertical axis is the False Positive Rate (which is represented as the vertical length of the x-axis). The resulting curve had an AUC score of 98.11%. A true-positive rate of 0.01 (also referred to as a one in a million likelihood) is plotted on the x-axis, while a false-positive rate of 0.99 (a one in a billion likelihood) is plotted on the y-axis, with the AUC score shown as 98.11%. We can see minor irregularities in the LR ROC curve by looking at Fig. 10 and 11, but the RF ROC curve is a smooth line that denotes an ideal model. The Table 1 represents the accuracy values of the LR and RF models with SMOTE. Here we can see that the accuracy of the LR is 97.31% which is very high. For a classification problem, accuracy is not only the comparative metric, other evaluation metrics should also be considered to compare two models. We can observe the evaluation metrics of both models in the Table 1.

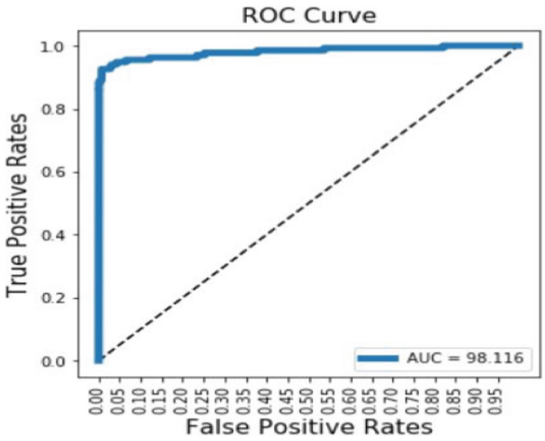

Fig. 10. ROC curve of LR

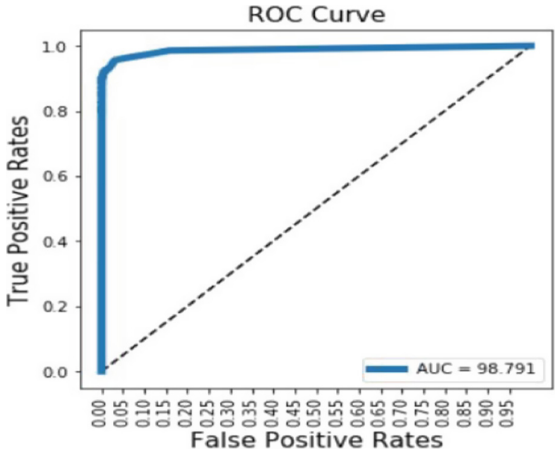

Fig. 11. ROC curve of RF

Table 1. Evaluation metrics with SMOTE (in %)

Models	Evaluation Metrics				
	Precision	Recall	F1 Score	AUC Score	Average PR Score
LR	95.22	92.64	9.89	98.11	78.32
RF	85.40	86.02	85.71	98.79	87.33

Table 2. Accuracy with SMOTE (in %)

Models	Accuracy
LR	97.31
RF	99.95

Table 2 exhibits the values of the evaluation metrics like precision, recall, F1 score, AUC score and average PR score for LR and RF with SMOTE. Here, the precision and F1 score values of LR are very low, which means that the performance of the model is not desired, whereas, all the evaluation metrics of RF are effective, which implies that it is the desired model. Compared to the LR model, the RF model gave the desired 99.95% accuracy and the performance metrics obtained were 85.40% precision, 86.02% recall, and 85.71% F1 score.

4 Conclusion

LR and RF algorithms are used in this paper to detect credit card fraud. The performance of the models is not correct when utilising an unbalanced dataset. To balance the imbalance dataset, use the SMOTE method. A more stable and generalised model can be achieved by balancing the dataset. The accuracy of the first LR model using SMOTE was 97.31%. Then, using SMOTE, an RF model was created that was 99.95% accurate. The assessment metrics for both models are compared, it is found that the RF with SMOTE is the best model for credit card fraud detection, with accuracy, precision, recall, and the weighted average of precision and recall of 99.95%, 85.40%, 86.02%, and 85.71%, respectively. RF with SMOTE has superior accuracy than LR with SMOTE. The proposed RF model can be used to anticipate credit card fraud in the E-Commerce industry, as credit card theft is on the rise. Using this technique, fraud transactions may be easily identified, and credit card fraud can be reduced in the future.

References

1. Sadgali, I., Nawal, S.A.E.L., Fouzia B.: Fraud detection in credit card transaction using machine learning techniques. In: 2019 1st International Conference on Smart Systems and Data Science (ICSSD), pp. 1–4. IEEE (2019)

2. Shirgave, S., Awati, C., More, R., Patil, S.: A review on credit card fraud detection using machine learning. Int. J. Sci. Technol. Res. **8**, 1217–1220 (2019)
3. Lakshmi, S.V.S.S., Kavilla, S.D.: Machine learning for credit card fraud detection system. Int. J. Appl. Eng. Res. **13**(24 Pt. 1), 16819–16824 (2018)
4. Bhanusri, A., Ratna Sree Valli, K., Jyothi, P., Varun Sai, G., Rohith Sai Subash, R.: Credit card fraud detection using machine learning algorithms. J. Res. Humanit. Soc. Sci. **8**, 4–11 (2020)
5. Dornadula, V.N., Geetha, S.: Credit card fraud detection using machine learning algorithms. Procedia Comput. Sci. **165**, 631–641 (2019)
6. Maniraj, S.P., Saini, A., Ahmed, S., Sarkar, S.: Credit card fraud detection using machine learning and data science. Int. J. Eng. Res. **8**(09) (2019)
7. Saragih, M.G., Chin, J., Setyawasih, R., Nguyen, P.T., Shankar, K.: Machine learning methods for analysis fraud credit card transaction. Int. J. Eng. Adv. Technol. **8**, 870–874 (2019)
8. Navanshu, et al.: Credit card fraud detection using machine learning models and collating machine learning models. Int. J. Pure Appl. Math. **118**(20), 825–838 (2018)
9. Zareapoor, M., Shamsolmoali, P.: Application of credit card fraud detection: based on bagging ensemble classifier. Procedia Comput. Sci. **48**(2015), 679–685 (2015)
10. Ramisetty, U.M., Chennupati, S.K.: Optimization of number of base station antennas in downlink massive MIMO and analysis of imperfect channel state information by perfection factor. Eng. Sci. Technol. Int. J. **23**(4), 851–858 (2020)
11. Ramisetty, U.M., Chennupati, S.K.: Performance analysis of multi user MIMO system with successive hybrid information and energy transfer beamformer. Wirel. Pers. Commun. **120**(1), 249–267 (2021). https://doi.org/10.1007/s11277-021-08450-y
12. Ramisetty, U.M., Chennupati, S.K., Venkata Nagesh Kumar, G.: Design of training sequences for multi user—MIMO with accurate channel estimation considering channel reliability under perfect channel state information using cuckoo optimization. J. Electr. Eng. Technol. **16**, 2743–2756 (2021)
13. Dheepa, V., Dhanapal, R.: Behavior based credit card fraud detection using support vector machines. ICTACT J. Soft Comput. **2**(4), 391–397 (2012)
14. More, R., Awati, C., Shirgave, S., Deshmukh, R., Patil, S.: Credit card fraud detection using supervised learning approach. Int. J. Sci. Technol. Res. **9**, 216–219 (2021)