ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 9, Issue 7, July 2022

Effects of Feature Selection with Machine Learning Algorithms in Detection of Credit Card Fraud

[1] Surbhi Bhardwaj, [2] Sonika Gupta

[1] [2] School of Computer Science and Engg, Shri Mata Vaishno Devi University, Katra, Jammu & Kashmir, India.
Corresponding Author Email: [1] 20mms016@smvdu.ac.in, [2] sonika.gupta@smvdu.ac.in

Abstract— As an effect of developments in e-commerce systems and communication technologies, Credit cards have become the most common mode of payment for purchases. The payments through the credit cards also involve the risk of credit card fraud such as application fraud, identity theft, lost/stolen card misuse, and phishing. These frauds lead to huge losses and require automatic and real-time fraud detection. Many studies have used Machine Learning (ML) techniques to detect fraudulent transactions. This study focuses on proposing a framework for the detection of credit card frauds by applying machine learning techniques like Random Forest (RF) and Naïve Bayes and testing the results on balanced and unbalanced datasets with and without performing the feature selection on the dataset. After comparing the results, it was discovered that Random Forest outperformed the Naïve Bayes on a balanced dataset with feature selection performed using Recursive Feature Elimination and Information Gain.

Index Terms— Machine Learning, Feature selection, SMOTE, Credit Card Fraud detection.

I. INTRODUCTION

The increased global online purchasing via the internet necessitates the use of Credit cards on a regular basis. Furthermore , the rapid and the total number of transactions made using a credit card (CCT) resulted in a significant increase in fraud cases . As a result , it is prudent to develop new methods and techniques for detecting these frauds [4] .Credit card fraud occurs when a fraudster obtains credit card information and uses it to make purchases without the owner's permission. Due to the widespread use of credit cards and the scarcity of reliable security systems, credit card fraud results in billion-dollar losses.

Because credit card companies are often reluctant to disclose such information, it's difficult to get a precise estimate of the losses [3].

As per the survey conducted by Business Today [17] , in FY21, there were 83,638 banking fraud cases in India, totaling Rs 1.38 lakh crore. According to the data provided by the RBI in response to an RTI request filed by India Today, only Rs 1,031.31 crore has been recovered thus far.To overcome these obstacles machine learning and data mining algorithms have been proposed such as deep belief networks , convolutional neural networks , recurrent neural networks , hidden markov model ,random forest , naive bayes etc.Rather than relying solely on traditional machine learning algorithms, features selected using a variety of feature selection techniques have had a significant impact on machine learning techniques' performance. Xuan [14] proposed CART (Classification and Regression Tree) based random forest for the classification of legitimate and fraudulent transactions . However , the random forest's performance improves after using the feature selection methods . Saheed [6] tested GA as a feature selection method along with the random forest , support vector machine , and naive bayes and discovered that random forest with GA outperformed the naive bayes and support vector machine . The GA for feature selection has been questioned for being overly complex in order to detect fraud with a high likelihood of over-fitting . This paper compares the results of random forest and naive bayes with different feature selection methods in detecting fraud as legitimate or fraudulent transactions, as well as the gaps or challenges identified in all of the papers reviewed.

The rest of the paper is organized as follows. Section II discusses the previous fraud detection systems and their effects , the results and the gaps identified in detection of credit card frauds. Section III consists of the proposed methodology of this research to analyze the effectiveness of feature selection with machine learning algorithms. In Section IV , all experimental results are presented that show the importance to train algorithms only on the relevant specific features. Section V summarizes and concludes the work.

II. REVIEW OF THE LITERATURE

The goal of fraud detection is to recognize whether a credit card transaction is legitimate or fraudulent , which is viewed as a classification problem. Credit card extortion can be identified with a good understanding of fraud detection advances.The summary of the reviewed papers is as follows.

Javad Forough [1] proposed an ensemble model which uses recurrent neural networks as base classifier and Feed forward neural network (FFNN) is used as voting mechanism after aggregation of different RNN classifiers results . A number of GRU and LSTM networks are used for recurrent networks that serve as base classifiers on various dataset samples, with the results being used to train the FFNN. In terms of both training and testing time, the ensemble approach based on GRU is more efficient than the one based on LSTM. This is because GRU has fewer parameters and gates than LSTM.

Xinwei Zhang [2] utilized HOBA (homogeneity-oriented behavior analysis) as a feature engineering method with a deep learning architecture as a fraud detection system . The features are selected based on the common characteristics by using transaction aggregation strategy . Out of CNN , DBN and RNN , DBN gives better F1-Score of 0.568 , Precision of 62.6% , Accuracy of 98.25% and AUC of 0.976.The findings also show that all data mining methods benefit from HOBA-based feature engineering when it comes to detecting fraudulent transactions.

Taha and Malebary [3] utilized an optimized lightGBM(light Gradient boosting machine) in which to tune the parameters of the light gradient boosting machine algorithm a Bayesian-based hyperparameter optimization technique is implemented. The most key features are chosen using the Information Gain approach, and the model's performance is evaluated using a 5-fold CV test.The Optimized light gradient boosting algorithm achieved the higher accuracy , AUC and F1-Score of 98% , 0.9094 and 0.5695 respectively.Even in unbalanced data sets, the P-R curve gives a complete picture of the classification's performance.

Rtyali and Enneya [4] suggested a hybrid anomaly detection approach that combines supervised and unsupervised detection using the machine learning techniques such as to extract the better prediction features use the SVM-RFE(Support Vector Machine-Recursive Feature Elimination) approach, the SMOTE technique for balancing an unbalanced dataset and the GridSearchCV approach was employed as a Hyper Parameter Optimization (HPO) by a Random Forest Classifier.The proposed model is denoted as RFC(HPO , RFE) , this hybrid method outperformed other state of the art methods of machine learning with accuracy of 99%,sensitivity of 95% and AUPR 0f 0.81.It's a reliable classifier model since it maintains a high level of accuracy regardless of data quantity.

Lucas [5] implemented automated feature engineering using a multi-perspective Hidden markov model . The model learns eight different HMMs using a combination of three binary perspectives: cardholder/ merchant, genuine/ fraudulent and amount/ timing. Finally, a set of eight HMM-based features will provide data on the validity and fraudulence of both terminal and cardholder histories. A Random Forest is trained for the classification of fraudulent and legitimate transactions based on the selected features.When HMM-based features are added to the existing transaction aggregation strategy, the precision–recall AUC of random forest classifiers improves consistently and significantly.

Yakub K. Saheed [6] used GA as a feature selection technique with Random Forest , SVM and Naive Bayes algorithms . On a German dataset RF with GA performed better with accuracy of 96.4 , recall of 96.4 and precision of 96.5.

Zhenchuan Li [7] employed deep neural networks with transaction aggregation strategy as feature selection technique while SMOTE is used for balancing the data collected from a financial company of china .The F1-Score of this model is 0.813, and the AUC PR is 0.825.

Priyanka Kumari [8] proposed a model with classifiers as bagging , voting and CART without applying any feature selection techniques . On the German dataset the results are concluded as CART gives better accuracy of 95.21% , precision and recall of 0.952.

Ugo Fiore [9] utilized generative adversarial networks for the classification of fraudulent and legitimate transactions without any feature selection strategy on European cardholders dataset.With this framework , sensitivity was improved at the expense of a little increase in false positives.

Pumsirirat & Yan [10] proposed a deep learning based model for the detection of fraudulent transactions . Two unsupervised learning methods of deep learning i.e autoencoders (AE) and restricted boltzmann machines (RBM) are employed in this model. AE used backpropagation to reconstruct the error. AE and RBM are two deep learning methods for detecting fraud in real time using normal transactions. The AUC score of AE is 0.9603 on a dataset of 284, 807 transactions, and the RBM-based AUC score is 0.9505.For larger datasets, it can be concluded that AE and RBM produce high AUC scores and accuracy.

Randhawa [11] utilized a total of fraud detection algorithms based on machine learning.The algorithms include everything from basic neural networks to deep learning models . In addition, for the creation of hybrid models, the AdaBoost and majority voting methods are used. The model's performance is assessed using a 10-fold cross validation approach. SVM outperformed all twelve algorithms with the highest MCC score of 0.813.Adaboost with SVM increased the fraud detection rate from 79.8% to 82.3% while the best rate for fraud detection was achieved by NN and NB at 78.8% in majority voting.

Sanaz Nami [12] designed a model with dynamic random forest (RF) and KNN for the classification of fraudulent and legitimate transactions without any feature selection method on a private bank dataset. It was shown that evaluating the resemblance of existing transactions in a cardholder's profile to test transactions could be utilized to detect payment card fraud successfully.

Xuan [14] employed random tree based RF and CART based RF for the classification , CART (Classification and Regression trees) based RF performed better in comparison of random tree based RF with the accuracy of 96.77% , recall of 95.27% and F-measure of 0.9601 but the precision is little worse,on a very large dataset of 30,000000 instances from an e-commerce company of china.

Kang Fu [15] proposed a model in which Convolutional neural network is used with a transaction aggregation strategy to select the predictive features.The model was executed on commercial bank data that was balanced using a cost based sampling measure . This proposed method outperforms other state-of-the-art methods when tested .

Table 1 . A summary of papers that compare existing credit card fraud detection systems.

Author(s)	Methods Used	Feature selection	Datasets	Balancing Techniques	Performance	Gaps
Forough J. & Momtazi S. (2021) [1]	Long short term Memory GRU Feed forward neural network 5-foldCV	Not done	European card dataset Brazilian dataset	Imbalanced data	The Ensemble of LSTM as base classifier performs better than GRU and GRU Ensemble.	Number of LSTM classifiers used increases the training and testing time.
Zhang, X., Han, Y., Xu, W., & Wang, Q. (2021) [3]	LightGBM Bayesian-based hyperparameter optimization Algorithm 5-fold CV	Information Gain (IG)	European dataset UCSD-FICO Data mining Contest 2009 Dataset	Imbalanced data	LightGBM has the highest AUC of 92.88 % after optimization.	Overfitting is a problem with light GBM. Datasets are imbalanced.
Rtayli N., & Enneya N. (2020) [4]	Random Forest GridSearchCV 10 fold CV	SVM-RFE	European data PaySim data	SMOTE	Dealing with large amounts of training data is easy. RFC is a reliable classifier in terms of noise and outliers.	A huge number of trees can slow down the process.
Lucas Y et al.(2020) [5]	Random forest	Hidden markov Model	Belgian credit cards data	Imbalanced data	RF can use sequential information for classification because of the HMM-based features.	Dataset is Imbalanced
Pumsirirat, A., & Yan, L. (2018) [10]	Autoencoder(AE) Restricted Boltzmann Machine (RBM)	Not done	German dataset Australian Dataset European dataset	Imbalanced data	For larger datasets, AE and RBM yield high AUC and accuracy. AUC's score of AE : 0.9603 AUC's score of RBM : 0.9505	Datasets are imbalanced
Randhawa K. et al (2018) [11]	Adaboost Majority Voting NB,RF,DT,GBT,RT,SVM,MLP,NN,LIR,LOR,DL,DS 10 fold CV	Not done	Financial Institution dataset from Malaysia	Imbalanced data	With adaboost, NB achieves 100% accuracy and an MCC score of 1.000. All models perform well in majority voting, with DT + GBT yielding an MCC score of 1.00.	Dataset is imbalanced Without adaboost and majority voting, linear regression gives a weak MCC score of 0.272.
Nami, S., & Shajari, M. (2018) [13]	Dynamic Random forest kNN	Not done	Private Bank Dataset	Imbalanced data	DRF produces a smaller number of trees than RF.	No data balancing

Bahnsen, A. C. et al (2016) [16]	Decision tree Logistic Regression Random Forest Bayes Minimum Risk(BMR) Cost sensitive Decision tree (CSDT)	Transaction aggregation strategy,Von Mises distribution	European card processing company	Imbalanced data	Out of all the algorithms, the CSDT algorithm performs the best	Dataset is imbalanced
Zhan, X. et al (2021) [2]	Deep Belief Networks (DBN) CNN RNN BPNN SVM RF	Homogeneity-oriented behavior analysis (HOBA) RFM	Commercial bank of China	Imbalanced data	DBN performs better than all algorithms with accuracy of 98.25% , AUC score of 0.976 and 51.96% of recall.	Can be computationally intensive as HOBA generates a larger number of variables set.
Fu, K., (2016) [15]	CNN	Aggregation Strategies	Commercial bank data	Cost Based sampling method	The CNN model is well-suited to training large amounts of data and includes a mechanism to prevent overfitting.	The training process can take a long time if the CNN has several layers.
Yakub K. Saheed et al.,(2020) [6]	Naive Bayes, Random forest SVM	Genetic algorithm	German Dataset	Imbalanced data	RF achieved the highest accuracy and sensitivity of 96.4% for both.	GA selects features iteratively and can be complex in larger datasets. Dataset is imbalanced
Lakshmi, S. V. S. S., & Kavilla, S. D. (2018) [18]	Logistic regression Decision tree Random forest	Not done	European bank data set	Oversampling Technique	Random forest performs better out of all three algorithms with accuracy of 95.5% .	Only accuracy is considered for evaluation of performances.

III. METHODOLOGY OF RESEARCH

The work presented in this research is focused on comparison of different results of Random Forest and Naive Bayes Algorithms when applied to European card holders dataset which is downloaded from Kaggle. This research work is mainly divided into three categories . Firstly, Random forest and Naive Bayes algorithms are trained on the European card holders dataset which was partitioned into two sets one is unbalanced dataset and the other is balanced dataset. The dataset balancing is done using SMOTE technique. Secondly , the Recursive feature Elimination and Information gain are applied on unbalanced dataset to extract specific set of features which are then used for the training of the algorithms. Lastly, the algorithms are trained on features selected from a balanced dataset after execution of Recursive Feature Elimination and Information gain.

Figure 1 Depicts the suggested system's framework as well as the methodology employed in this study.

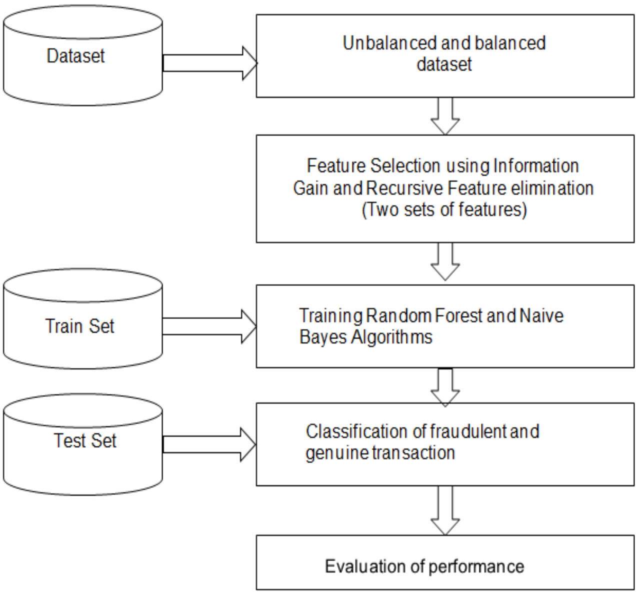
Dataset
Unbalanced and balanced dataset
Feature Selection using Information Gain and Recursive Feature elimination (Two sets of features)
Train Set
Training Random Forest and Naive Bayes Algorithms
Test Set
Classification of fraudulent and genuine transaction
Evaluation of performance

Figure1: The proposed framework for credit card fraud model

A. Dataset

European Card holders dataset from kaggle is utilized in this research [19]. This dataset has 31 features in total and out of 284,807 transactions only 492 are fraudulent. Fraudulent transactions account for 0.172 percent of all transactions, indicating that the dataset is severely imbalanced.This dataset is balanced using SMOTE technique.

B. Feature Selection

Recursive Feature elimination and Information gain methods are used for the selection of most specific features to train the random forest and Naive Bayes models. Out of 31 total number of features a set of 14 features is extracted by both these methods.

C. Evaluation Metrics

In this research work , Positives class (P) denotes the number of fraudulent transactions whereas negatives (N) denotes the number of authentic transactions.

TP –True Positive, TN - True Negative, FP – False Positive, FN – False Negative.

1. Accuracy: It represents out of all classes i.e positive or negative , how many predictions are correct.

   $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Recall : It represents what proportion of actual positive classes was correctly identified by the model.

   $$\text{Recall} \;=\frac{TP}{TP + FN}$$

3. AUC - ROC Score : The ROC curve is a plot between True Positive Rate (TPR) and the False Positive Rate (FPR). The AUC Score ranges between 0 and 1.

IV. EXPERIMENTAL RESULTS

In this research , European card holders' dataset is employed for the experiments . Initially the dataset comprises 31 features. Recursive Feature Elimination and Information gain are used to extract specific features for the training of random forest and naive bayes.

A. Results analysis without Feature Selection

Table 2 Consists of results of random forest and naive bayes when feature selection techniques are not applied on balanced and unbalanced datasets.

Table : 2

Evaluation Metrics	RF on Balanced dataset	RF on UnBalanced dataset	NB on Balanced dataset	NB on UnBalanced dataset
Accuracy	99.9	99.9	97.6	97.8
Recall	82.0	83.8	88.2	84.5
ROC_AUC Score	91.0	91.9	92.9	91.1

B. Results analysis with Recursive Feature Elimination (RFE)

Table 3 Consists of results of random forest and naive bayes when features are selected using RFE.

Table 3

Evaluation Metrics	RF on Balanced dataset with RFE	RF on UnBalanced dataset with RFE	NB on Balanced dataset with RFE	NB on UnBalanced dataset with RFE
Accuracy	99.9	99.9	97.8	97.7
Recall	88.9	78.6	88.9	85.2
ROC_AUC Score	94.4	89.3	93.4	91.5

C. Results analysis with Information Gain (IG)

Table 4 Consists of results of random forest and naive bayes when features are selected using the Information gain method.

Table 4

Evaluation Metrics	RF on Balanced dataset with IG	RF on UnBalanced dataset with IG	NB on Balanced dataset with IG	NB on UnBalanced dataset with IG
Accuracy	99.9	99.9	98.0	98.4
Recall	87.5	81.6	88.9	87.5
ROC_AUC Score	93.7	90.8	93.5	92.9

In results of table 2 , table 3 and table 4 it is shown that when Random forest with Recursive feature Elimination (RFE) used on a balanced dataset achieved a recall of 88.9 % and ROC_AUC Score of 94.4 % which are the highest result scores out of all experiments done. Naive Bayes followed by Random Forest achieved a recall of 88.9% and ROC_AUC score of 93.5 % when Information Gain was applied for feature selection on a balanced dataset.

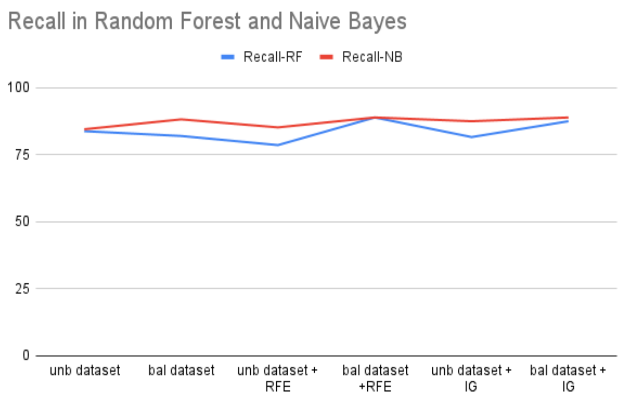

Figure 2 : Recall comparison for Random Forest and Naive Bayes when applied to balanced and unbalanced datasets with and without the use of Recursive Feature Elimination and Information Gain.

V. CONCLUSION

Random forest and Naive Bayes take less time in training when applied to larger datasets . The Findings of the experimental result have shown that when datasets are balanced and a relevant set of features are selected using feature selection techniques, random forest performs equally well compared to the deep learning models. The future Scope of this research is to test results for overfitting problems and to enhance the results if overfitting problems exist. The results of Random Forest and Naive Bayes can be tested for more bigger size datasets.

REFERENCES

[1] Forough, J., & Momtazi, S. (2021). Ensemble of deep sequential models for credit card fraud detection. Applied Soft Computing, 99, 106883.

[2] Zhang, X., Han, Y., Xu, W., & Wang, Q. (2021). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. Information Sciences, 557, 302-316.

[3] Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access, 8, 25579-25587.

[4] Rtayli, N., & Enneya, N. (2020). Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization. Journal of Information Security and Applications, 55, 102596.

[5] Lucas, Y., Portier, P. E., Laporte, L., He-Guelton, L., Caelen, O., Granitzer, M., & Calabretto, S. (2020). Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs. Future Generation Computer Systems, 102, 393-402.

[6] Saheed, Y. K., Hambali, M. A., Arowolo, M. O., & Olasupo, Y. A. (2020, November). Application of GA Feature Selection on Naive Bayes, Random Forest and SVM for Credit Card Fraud Detection. In 2020 International Conference on Decision Aid Sciences and Application (DASA) (pp. 1091-1097). IEEE

[7] Li, Z., Liu, G., & Jiang, C. (2020). Deep representation learning with full center loss for credit card fraud detection. IEEE Transactions on Computational Social Systems, 7(2), 569-579

[8] Kumari, P., & Mishra, S. P. (2019). Analysis of credit card fraud detection using fusion classifiers. In Computational Intelligence in Data Mining (pp. 111-122). Springer, Singapore.

[9] Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences, 479, 448-455.

[10] Pumsirirat, A., & Yan, L. (2018). Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. International Journal of advanced computer science and applications, 9(1), 18-25.

[11] Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. IEEE access, 6, 14277-14284.

[12] Carcillo, F., Le Borgne, Y. A., Caelen, O., & Bontempi, G. (2018). Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization. International Journal of Data Science and Analytics, 5(4), 285-300.

[13] Nami, S., & Shajari, M. (2018). Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors. Expert Systems with Applications, 110, 381-392.

[14] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & 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.

[15] Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016, October). Credit card fraud detection using convolutional neural networks. In International conference on neural information processing (pp. 483-490). Springer, Cham.

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

[17] Business Today the business magazine India (Dec 15, 2021) available https://www.businesstoday.in/union-budget-2022/banking/story/india-saw-229-banking-frauds-per-day-in-fy21-less-than-1-amount-recovered-315685-2021-12-15.

[18] Lakshmi, S. V. S. S., & Kavilla, S. D. (2018). Machine learning for credit card fraud detection system. International Journal of Applied Engineering Research, 13(24), 16819-16824.

[19] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud