ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2023.0434
SPECIAL ISSUE ON INNOVATIONS AND TECHNOLOGY IN FINTECH (ITFT-2023)

ENHANCING CREDIT CARD FRAUD DETECTION IN FINANCIAL TRANSACTIONS THROUGH IMPROVED RANDOM FOREST ALGORITHM

B. Sowmiya
Department of Computer Science, S.I.V.E.T. College, India

Abstract

Credit card Fraud detection is a critical task in various industries, including finance and e-commerce, where identifying fraudulent activities can help prevent financial losses and protect users. It begins by combining two datasets containing fraudulent and non-fraudulent transactions to create a comprehensive dataset for analysis. Data is preprocessed by removing unnecessary features, calculating distance metrics, and generating new variables to capture temporal patterns and transaction history. Multicollinearity issues are addressed through feature selection. Improved Random Forest (RF) algorithm is used to improve fraud detection. The experimental results indicate that the improved Random Forest algorithm achieves commendable accuracy in fraud detection. The proposed model achieves 99.87% training accuracy and 99.41% testing accuracy. The Model's performance is evaluated by measuring precision, recall, F1-score and support. Our research emphasizes the importance of considering improved algorithms to achieve better results. The findings provide valuable insights for organizations aiming to enhance their fraud detection capabilities and make informed decisions to protect their systems and users.

Keywords:

Credit Card, Fraud detection, Random Forest, Classification, Accuracy, Precision, Recall, and F1 Score

1. INTRODUCTION

Fraud detection is an important task in various industries, especially in the financial sector. The ability to identify fraudulent activities can help prevent financial losses and maintain the integrity of transactions. In this research, we explore the application of machine learning techniques for fraud detection using a dataset containing transaction information.

With increase in credit card usage, fraudulent activities have become a significant concern for card holders. Among inner card fraud and external card fraud, the latter is accountable for the majority of credit card frauds [1]. Through many means fraudulent activities continue to happen. It's quite interesting that card owners are affected the least compared to merchant, because the former have limitations than the later [2]. Credit card fraud may be attempted online or offline. Through virtual mode, illegal fraud activities happen without the knowledge of users. User information such as card number, account number and other details are stolen to perform fraudulent activities [3].

A lot of studies have been conducted on credit card fraud detection. One such research uses machine learning techniques to develop an effective model that can accurately identify fraudulent transactions in credit card data. Various machine learning algorithms, such as Support Vector Machines (SVM), are employed to classify transactions as legitimate or fraudulent based on historical data. In the study on credit card fraud detection using machine learning, various algorithms were employed, including logistic regression, decision trees, and support vector machines [4]. Machine learning approaches analyses behavioral patterns of credit card transactions and enhances security in the financial industry [5]. Next advancement in technology: Neural Networks (NN), which is inspired by human brain. Artificial neural networks are a solution for detecting credit card fraud. They are capable of learning complex patterns and relationships from large datasets, making them suitable for fraud detection tasks. It explains that neural networks simulate the human learning process and consist of artificial neurons organized in layers [6]-[7]. With advancements in research, machine learning models and artificial neural networks are combined to tackle this problem [8]. Some researchers use variants of Machine learning algorithms for fraud detection [9]. Deep learning is a subset of machine learning. It is widely used in many applications including fraud detection [10]. Unlike machine learning algorithms, feature selection is automated in deep learning [11]-[12].

The objective of the study is to develop a system for credit card fraud detection. Section 2 explores the existing work done by various researchers. Section 3 presents the implementation of proposed system and detailed technical explanation of proposed improved Random Forest machine learning algorithm. Section 4 elaborates the results and discussion section with enough visualization. Section 5 concludes the paper with futuristic work.

2. LITERATURE REVIEW

The authors highlight the increasing incidence of credit card fraud and the need for effective detection mechanisms. They propose the random forest algorithm as a potential solution due to its ability to handle large datasets and deal with the imbalanced nature of fraud detection datasets. The research suggests that the random forest algorithm can be a valuable tool in credit card fraud detection and provides insights into the application of this algorithm and its potential to enhance the security of credit card transactions by accurately identifying fraudulent activities [3].

The models were trained and evaluated on a large dataset of credit card transactions. The results demonstrated the effectiveness of machine learning in detecting fraudulent activities, achieving high accuracy rates ranging from 90% to 95%. The study aims to improve the security and integrity of credit card transactions by leveraging the power of machine learning for fraud detection. The study highlights the potential of machine learning techniques in enhancing fraud detection systems and improving overall security in financial industry [4].

The paper presents a behavior-based approach for credit card fraud detection using Support Vector Machines (SVM). By analyzing the behavioral patterns of credit card transactions, SVM models are trained to identify fraudulent activities. The results demonstrate the effectiveness of the proposed method, achieving high accuracy rates of over 95%. The study highlights the

potential of behavior-based approaches and SVMs in detecting credit card fraud, contributing to enhanced security in the financial industry [5]. Stolen or Lost Cards, Counterfeit Cards, Card-Not-Present (CNP) Fraud, Identity Theft, Skimming, Account Takeover, Phishing and Social Engineering, Friendly Fraud are the fraud techniques addressed by the authors. NN makes use of patterns to authorize transactions. Back Propagation is used for training purpose. Genetic Algorithm and Neural Network (GANN) combination is discussed as an innovative ideology for fraud detection [6].

The authors describe the experimental setup for credit card fraud detection using neural networks. Research utilizes perceptron model, a basic type of artificial neural network, and the multilayer perceptron, which is a more complex model capable of delivering outputs with more than two classes. The dataset used for training and testing consists of transaction details of 20000 active credit card holders over the past six months. The authors explain how they categorized transactions as fraudulent or legitimate based on various factors such as spending areas, number of transactions, average monthly balance, and total purchase amount. They utilize Neuroph, a lightweight Java Neural Network Framework, to develop and train the neural network [7]. Preprocessing techniques, such as feature selection and correlation analysis, are applied to the dataset. Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), RF and Artificial neural network (ANN) are utilized. Two approaches are used to address class imbalance: resampling the dataset and applying class weights to the classifiers. From their experiment, RF outperforms with more than 96% accuracy [8].

The study focuses on supervised learning techniques and compares the performance of Decision Tree, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbor (KNN) algorithms for credit card fraud detection. Study confirms that SVM outperform DT, KNN and RF. [9]. Authors propose the use of a deep learning model based on an auto-encoder (AE) and a restricted Boltzmann machine (RBM) to detect anomalies in normal transaction patterns. The implementation of the AE and RBM models is done using the TensorFlow library [10].

3. PROPOSED WORKFLOW

The proposed system creates an enhanced mechanism for credit card fraud detection in finance transactions. This section provides an in-depth explanation of the proposed system.

3.1 DATA ACQUISITION

The data is acquired from public dataset, Kaggle. The dataset contains two folders for training and testing. https://www.kaggle.com/datasets/kartik2112/fraud-detection.

3.2 PREPROCESSING

Preprocessing improves accuracy by eliminating noise and potential distractions from unnecessary data. In the data preprocessing step, four columns were removed from the dataset. These columns include 'Unnamed: 0', 'trans_num', 'first', 'last', 'street', and 'city'. The removal of these columns was deemed necessary either due to their redundant nature or because they contained sensitive or irrelevant information for the analysis or modeling task. This preprocessing step helps streamline the dataset for further improving accuracy and efficiency of fraud detection model.

3.3 MODEL ARCHITECTURE

The proposed architecture is depicted in Fig.1.

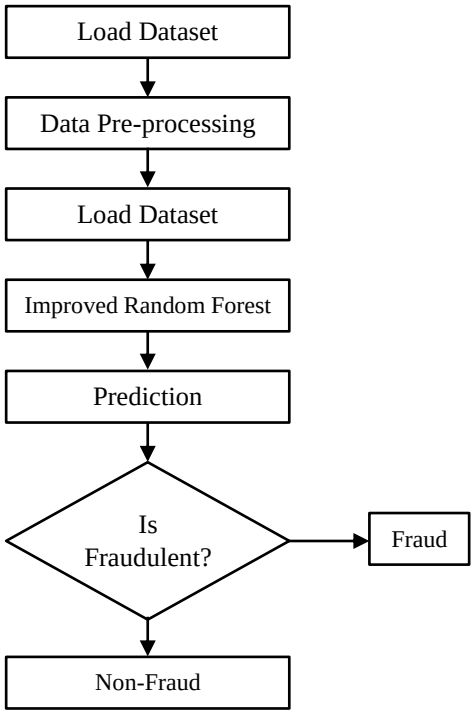

Fig.1. Proposed workflow

3.3.1 Proposed Random Forest Model:

We propose an improved RF model to identify fraud detection and the steps are explained below.

Step 1: Load the dataset.

Step 2: Preprocess the dataset by dropping irrelevant columns, calculating distance between locations, calculating sum of transactions in past 30 days.

Step 3: Data is split into train and test set.

Step 4: Model is trained and evaluated by improved RF classifier.

Step 5: Model makes predictions on test dataset.

Step 6: Accuracy score of the model is calculated.

3.4 CLASSIFICATION

It involves classifying data into different classes. This research involves binary classification, where the improved RF model is trained on a dataset, which is later used to predict the class. Dataset is divided into train and test set in 80:20 ratios.

Step 1: Data is split into test and train dataset.

Step 2: The model is trained by improved RF classifier.

Step 3: Data is trained on training data using ‘fit’ method.

Step 4: Model is predicted using ‘predict’ method.

Step 5: Accuracy of classification is predicted using ‘accuracy_score’ method.

3.5 IMPROVED RANDOM FOREST CLASSIFICATION ALGORITHM

RF algorithm has the ability to handle complex and high-dimensional data. Multiple decision trees are constructed using random subsets of features and samples. Each tree is trained to classify transactions as fraudulent or non-fraudulent based on the input features. During the prediction phase, each tree in the RF model independently classifies a new transaction as either fraudulent or non-fraudulent. The final classification is determined by aggregating the results from all the trees through voting or averaging.

$$RF(x) = \Sigma(T_i(x)) / N \quad (1)$$

where, $RF(x)$ represents the final prediction for a given input transaction x, $T_i(x)$ represents the prediction of the i_{th} decision tree in the random forest for transaction x and N represents the total number of decision trees in the random forest.

3.6 IMPROVED RANDOM FOREST

The Random Forest classifier in this code is created with optimized parameters, including n_estimators, max_features, max_depth, min_samples_split, and min_samples_leaf.

Feature selection is performed in this code to identify and select the most important features from the dataset, allowing the model to focus on the most informative attributes and improve its overall performance.

Out-of-bag error provides an estimate of the model's performance on unseen data. It enables the assessment of generalization ability and detection of overfitting or underfitting issues without the need for a separate validation set.

Cross-validation is performed to obtain a more reliable estimate of the model's performance by assessing its stability and generalization across different subsets of the training data. It leads to a more robust evaluation of the model's effectiveness in fraud detection.

4. RESULTS AND DISCUSSION

The proposed system utilized Intel(R) Core(TM) i3-5005U CPU with 4GB of RAM and a 64-bit processor. The software environments are Anaconda 1.9.0 and Python 3.9.7.

The correlation coefficient measures the strength and direction of linear relationship between two variables. A positive value indicates positive correlation and negative value indicates negative correlation. Correlation matrix is a square matrix that shows the pair wise correlations between variables in a dataset. The Fig.2 displays correlation matrix as heatmap, providing visual summary of correlations, with annotation values.

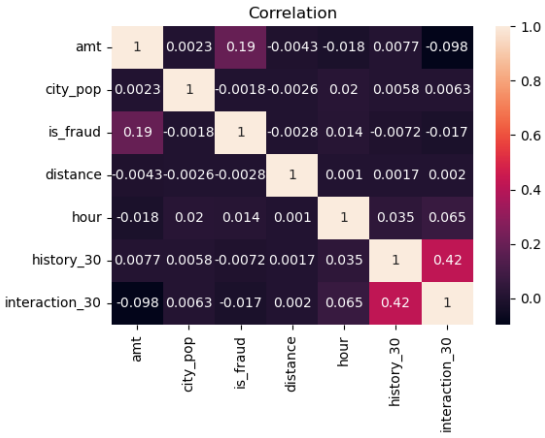

Fig.1. Correlation matrix

Confusion matrix evaluates the model's performance by comparing actual target values with the predicted values. Evaluation metrics like precision, recall, support, and F1-score to assess the efficiency of the classification system. Confusion matrix estimates are explained in Table.1.

Table.1. Key terms in Confusion matrix

Outcome	Explanation
TP	Correctly predicted positive values
TN	Correctly predicted negative values
FP	Incorrectly predicted as positive values
FN	Incorrectly predicted as negative values

The Eq.(2)-Eq.(4) show formula for precision, recall and F1.

$$\text{Recall} = TP/(TP+FN) \quad (2)$$

$$\text{Precision} = TP/(TP+FP) \quad (3)$$

$$\text{f1-score} = 2*((\text{precision}*\text{recall})/(\text{precision}+\text{recall})) \quad (4)$$

The Fig.3 and Fig.4 depicts the confusion matrix for Random Forest and improved Random Forest algorithm.

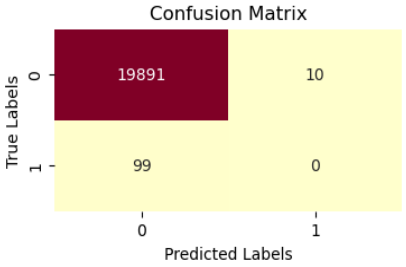

Fig.2. Confusion matrix - Random Forest algorithm

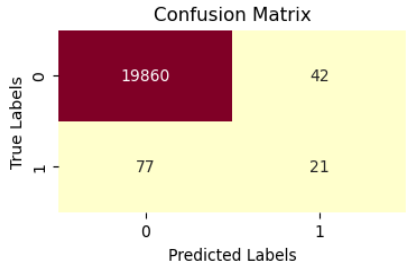

Fig.3. Confusion matrix – Improved RF algorithm

The Fig.5 represents ROC AUC curve for proposed model.

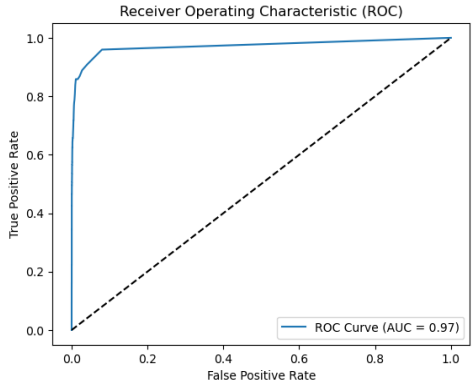

Fig.4. ROC-AUC curve of improved RF model

The Table.2 shows classification report for the proposed model. Precision, recall, F1-score and support are the parameters considered.

Table.2. Classification report

	Precision	Recall	f1-score	Support
0	1.00	1.00	1.00	19902
1	0.70	0.60	0.65	98
Accuracy			0.995	20000
Macro avg	0.85	0.80	0.65	20000
Weighted avg	0.99	0.995	0.99	20000

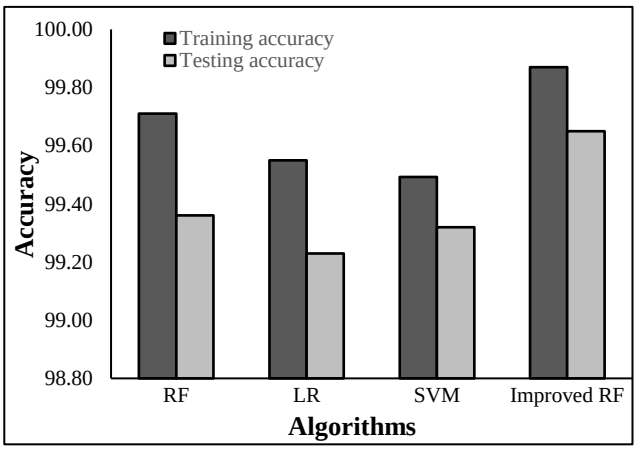

Fig.6. Accuracy of proposed method with other methods

The proposed method is tested against RF, LR and SVM. Experimental results indicate that improved RF outperforms other three classifiers in terms of training and testing accuracy. The Fig.6 depicts the accuracy of proposed method with other machine learning algorithms.

Aburbeian et al. [13] developed a similar system and the comparative results are summarized as follows.

- An accuracy of 98% for predicting transactions, with Class 0 (non-fraud transactions): Precision: 100%, Recall: 96%, F1-score: 98% and Class 1 (fraud transactions): Precision: 96%, Recall: 100%, F1-score: 98%. Results clearly indicate that the model performed exceptionally well for both classes, with minimal false positives and false negatives.
- The results of the proposed system are compared with [13], where our classifier gained overall accuracy of 99.87% for training and 99.41% for testing. An accuracy of 100% for predicting transactions, with Class 0 (non-fraud transactions): Precision: 100%, Recall: 96%, F1-score: 98% and Class 1 (fraud transactions): Precision: 70%, Recall: 60%, F1-score: 65%.
- The results clearly indicate that the proposed model performed exceptionally well for class 0, with minimal false positives and false negatives and slightly lower accuracy of 98.47% for Class 1, which needs to be improved further.

5. CONCLUSION

Early detection of fraudulent transactions allows timely action to be taken at an early stage which minimizes potential financial losses. Analyzing patterns and trends in fraudulent transactions can provide valuable insights into emerging fraud techniques and vulnerabilities. This information can be used to improve fraud prevention strategies and stay ahead of evolving fraud threats. Fraud detection is an ongoing process that requires constant adaptation and improvement to stay effective in the face of ever-evolving fraud techniques. The proposed system achieves 99.41% testing accuracy and 99.87% training accuracy. The limitations in this study are fraud detection done solely on previous month transaction. Combining historical data with additional features and more advanced fraud detection techniques, such as anomaly detection or network analysis, can enhance the overall fraud detection capability and the same will be futuristic work. Future studies will be conducted with cutting edge technologies like Convolutional Neural Network and Deep learning technologies.

REFERENCES

[1] A. Shen and Y. Deng, “Application of Classification Models on Credit Card Fraud Detection”, *Proceedings of International Conference on Service Systems and Service Management*, pp. 1-4, 2007.

[2] J.T. Quah and M. Sriganesh, “Real-Time Credit Card Fraud Detection using Computational Intelligence”, *Expert Systems with Applications*, Vol. 35, No. 4, pp. 1721-1732, 2008.

[3] M.S. Kumar, E.S. Keerthika and E. Aswini, “Credit Card Fraud Detection using Random Forest Algorithm”, *Proceedings of International Conference on Computing and Communications Technologies*, pp. 149-153, 2019.

[4] P. Tiwari and A.K. Singh, "Credit Card Fraud Detection using Machine Learning: A Study", *Proceedings of International Conference on Artificial Intelligence*, pp. 1-10, 2021.

[5] V. Dheepa and R. Dhanapal, "Behavior based Credit Card Fraud Detection using Support Vector Machines", *ICTACT Journal on Soft Computing*, Vol. 2, No. 4, pp. 391-397, 2012.

[6] R. Patidar and L. Sharma, "Credit Card Fraud Detection using Neural Network", *International Journal of Soft Computing and Engineering*, Vol. 1, pp. 32-38, 2011.

[7] D. Murli, D. Jog and S. Nath, "Credit Card Fraud Detection using Neural Networks", *International Journal of Students' Research in Technology and Management*, Vol. 2, No. 2, pp. 84-88, 2015.

[8] A. Sahu and M.K. Gourisaria, "A Dual Approach for Credit Card Fraud Detection using Neural Network and Data Mining Techniques", *Proceedings of IEEE International Conference on India Council*, pp. 1-7, 2020.

[9] I. Sadgali, N. Sael and F. Benabbou, "Fraud Detection in Credit Card Transaction using Neural Networks", *Proceedings of International Conference on Smart City Applications*, pp. 1-4, 2019.

[10] A. Pumsirirat and Y. Liu, "Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine", *International Journal of Advanced Computer Science and Applications*, Vol. 9, No. 1, pp. 1-13, 2018.

[11] X. Zhang and Q. Wang, "HOBA: A Novel Feature Engineering Methodology for Credit Card Fraud Detection with a Deep Learning Architecture", *Information Sciences*, Vol. 557, pp. 302-316, 2021.

[12] O. Voican, "Credit Card Fraud Detection using Deep Learning Techniques", *Informatica Economica*, Vol. 25, No. 1, pp. 1-12, 2021.

[13] A.M. Aburbeian and H.I. Ashqar, "Credit Card Fraud Detection using Enhanced Random Forest Classifier for Imbalanced Data", *Proceedings of International Conference on Advances in Computing Research*, pp. 605-616, 2023.