International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Credit Card Fraud Detection Using Machine Learning

Mr. NOOR AHAMED J[1] M.C.A., M.Phil., Veera Pragadesh.M[2]

1Assistant Professor (SG), Department of Computer Applications, Nehru college of management,

Coimbatore, Tamil Nadu, India.

jnamca@gmail.com

2II MCA, Department of Computer Applications, Nehru college of management,

Coimbatore, Tamil Nadu, India.

ABSTRACT

Credit card fraud poses a significant threat to financial institutions and cardholders, resulting in substantial financial losses and compromised security. This study focuses on the development of effective fraud detection systems using machine learning algorithms to mitigate these risks. Leveraging a comprehensive dataset containing transactional information, this research explores the application of various machine learning techniques to identify and prevent fraudulent credit card transactions. Credit card fraud detection is a critical task for financial institutions and merchants, as it can result in significant financial losses and damage to customer trust. Machine learning classifiers offer a promising solution for detecting fraudulent transactions in real-time. This project aims to develop a credit card fraud detection system using machine learning classifiers, which can accurately identify fraudulent transactions while minimizing false positives. The project involves collecting and preprocessing a large dataset of credit card transactions, training several machine learning classifiers, and selecting the best performing model. The final model will then be tested on a new dataset to evaluate its accuracy, precision, and recall. The resulting system can be integrated into financial institutions and merchants' existing fraud detection systems to enhance their ability to detect and prevent fraudulent transactions. This project has the potential to save millions of dollars in financial losses due to credit card fraud and increase customer confidence in the security of their financial transactions. In this project we can implement the framework to study the multiple classifiers such as Random Forest (RF), Linear Regression (LR), Decision tree classifier (DT) and Support Vector Machine algorithm (SVM) in credit card datasets that are collected from KAGGLE source and implement in Python framework.

1. INTRODUCTION

1.1 DATA MINING

Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system.

Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.

1.2 FOUNDATION OF DATA MINING:

Data Mining is the process of posing queries to large amounts of data sources and extracting patterns and trends using statistical and machine learning techniques. It integrates various technologies including database management, statistics and machine learning. Data mining has applications in numerous disciplines including medical, financial, defense and intelligence. Data mining tasks include classification, clustering, making associations and anomaly detection. For example, data mining can extract various associations between people, places or words. During recent years there have been many developments in data mining. The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as "knowledge discovery in databases," the term "data mining" wasn't coined until the 1990s. But its foundation comprises three intertwined scientific disciplines: statistics (the numeric study of data relationships), artificial intelligence (human-like intelligence displayed by software and/or machines) and machine learning (algorithms that can learn from data to make predictions). What was old is new again, as data mining technology keeps evolving to keep pace with the limitless potential of big data and affordable computing power. Various data mining techniques have been developed. These include techniques for extracting associations, neural networks, inductive logic programming, decision trees, fuzzy logic and rough sets. Furthermore, data mining has gone beyond mining relational databases to mining text and multimedia data. Also, data mining is being applied to areas such as information security and intrusion detection. While there have been many practical developments, we still have major challenges. One of the most important challenges is scalability. If data mining is to be useful we need to mine very large databases. Therefore, it is critical that we need to understand the limitations of the data mining algorithms. To understand the limitations, we need to study the foundations of data mining. We need to explore the time and space complexity of the algorithms. There are techniques such as inductive logic programming and rough sets that have underpinnings in logic and mathematics. One needs to explore these techniques for data mining and examine the computational complexity aspects. We also need to understand the complexity of the various search algorithms being used for market basket analysis.

2. RELATED WORKS

Sahu, Aanchal et al [1] Build new models to detect fraudulent credit card transactions using five classifiers to find out the best fit classifier for the situation. The dataset contains the details of some European credit card holders recorded in September 2013. The dataset comprises the transaction information for two days, which has 492 fraud transactions out of 284,807 transactions. For security purposes, the features of the dataset are not revealed. Instead, the PCA values of the features are given. A PCA is a technique to get a low dimensional structure out of a potential high dimensional dataset. It includes the extraction of q eigenvectors for q input distribution [10]. It is one of the most used algorithm for dimensionality reduction. The basis vectors are known as principal components. The dataset contains a total of 31 columns of which, 28 are PCA components named as V1, V2….V28. Moreover, the time and amount of the money transaction has also been provided. The target variable classifies a transaction as 0 for valid transaction and 1 for fraudulent transaction. Here apply classifiers of logistic regression (LR), support vector machine (SVM), decision tree (DT), random forest (RF) and artificial neural network (ANN) on two different approaches used on the same data. In both the approaches, the main goal was to curb the problem of data imbalance (since number of fraudulent cases are scarce in comparison to normal behaviour). In the first approach we resampled the minority class to a higher number close to the number of samples of normal class, while, in the second approach we use cost-based methods such as applying weights on classes so that the minority class has a higher impact on the loss (error) calculation for the models. After thorough experimentation, we notice that RF outperforms all other classifiers on an average on both the approaches.

Panda, Agyan et al [2], Implement credit card fraud detection methods based on data mining. Classic data mining algorithms aren't directly applicable to our topic because it's handled as a classification challenge. As a result, a different technique is employed, which involves the employment of general-purpose meta heuristics like genetic algorithms. The purpose of this study is to create a credit card fraud detection system based on genetic algorithms. Genetic algorithms are a form of evolutionary algorithm that tries to continuously improve solutions. When a card is duplicated, stolen, or lost by fraudsters, it is usually utilized until the available limit is exhausted. As a result, rather than focusing on the quantity of correctly classified transactions, a strategy that reduces the overall allowed limit on fraud-prone cards takes precedence. Its goal is to reduce false alerts by utilizing a genetic algorithm to optimize a set of interval-valued parameter.

This concept is difficult to put into practice in practice since it necessitates the collaboration of banks, which are unwilling to exchange information owing to market competition, as well as for legal concerns and the protection of their users' data. As a consequence, we searched up some reference publications that used comparable methods and gathered data. As stated in one of these reference papers: Credit card fraud is unquestionably a kind of criminal deception. This article evaluated current results in this subject and outlined the most prevalent types of fraud, as well as how to identify them. This study also goes into great depth on how machine learning may be used to improve fraud detection outcomes. Pseudo code, explanation its implementation and experimentation results.

Aziz, Amir, And Hamid Ghous et al [3] intend to contribute to a first step in combating false news, commonly known as stance identification, in which the task is to determine the stance of a claim with relation to another piece of material. Our tests are based on the setting for the inaugural Fake News Challenge (FNC). In FNC1, the assertion appears in the form of a headline, whereas the other piece of material is an article body. This stage may appear, and is, a long way from automatically evaluating the truthfulness of a piece of material against some type of ground truth. However, the issue rests precisely in the notion of truth and the fact that it is vulnerable to prejudice. Furthermore, and in part because of this, annotated corpora for training and experimental assessment are difficult to find and are not always publicly available (as in the case of fact checker archives). We argue that determining whether a piece of content is related or unrelated to another piece of content (for example, headline vs. article body) is an important first step, which could be described as click bait detection. Following the FNC1 setup, the further classification of related pieces of content into more fine-grained classes provides valuable information once the "truth" (in the form of a collection of facts) has been established, allowing specific pieces of content to be classified as "fake" or, rather, "false". Since this definitive, resolving collection of facts is usually difficult to come by, the challenge of stance detection can be used to combine the outcome with credibility or reputation scores of news outlets, where several high-credibility outlets disagreeing with a particular piece of content point towards a false claim. Stance detection may also be useful for identifying political bias: if writers on the same end of the political spectrum are more likely to agree with each other, the (political) preference of one author can be induced once the preference of the other author is known.

Shah, Ankit, And Akash Mehta et al [4] I Implement six widely used machine learning techniques for credit card fraud detection. For each machine learning technique, a confusion matrix is prepared for performance analysis of the algorithm. Their efficacy is analyzed based on the parameters such as accuracy, precision, recall, specificity, misclassification, and F1 score. Results of fraud detection techniques also depend on a type of dataset. Some techniques give high accuracy but training them is very expensive. With small data sets, some techniques give excellent results, but they do not apply to large datasets. With sampled and pre-processed data, some techniques give better results whereas some techniques give better accuracies with raw unstapled data. It is also important to note that the outlier class of modeling can be senseless and unproductive in solving the problem of anomaly detection. There is a need to focus on the structure of the normal data and its distribution. Credit card fraud detection system should be capable of detecting fraud in the transit process and to identify fraud precisely and wrong classifications should have to be minimum. There is a need for a technology that should be able to detect fraudulent transactions when it is occurring. The best result cannot be achieved by applying one machine learning technique. Therefore, to achieve better performance for credit card fraud detection the integration of multiple techniques may be used. In the future, researchers can compute the computational complexity and execution time of various algorithms to suggest the best possible solution.

Muttipati, Appala Srinuvasu et al [5] This proposed work going to address the problem of an imbalanced dataset. SMOTE sampling technique is used to convert the imbalanced dataset to a balanced binary dataset. The credit card dataset which we have chosen for reference dataset consists of an error of 0.172 percent. It gives a meaning that, the referenced dataset contains 0.172 percentage of no genuine transactions. This infers that, the dataset is uneven and is prejudiced towards genuine transaction. Because of the bias the network is unable to identify and could give a correct prediction of the error. This problem can be solved by using 2 techniques, i) under-sampling ii) over-sampling techniques to condense the partiality for accurate results. Under-sampling technique, balances the dataset by basing on the non-bias class i.e. fraudulent Transactions. By adjusting, total of genuine transactions on equality with fraudulent transactions by removing the excess genuine values from the data. For example, suppose there is an 100 observations, then 7 fraudulent values give a 7% error. Similar to that we compute the total of genuine transactions for 492 fraudulent ones, by removing the excess. This produces data with 3% error, it become because easier to detection process. By utilizing this method it results to loss of information. Over-sampling is another technique that is utilized for imbalanced to

balanced data. Here, the occurring of bias is due to a replica of information in terms of the recurring rows but not for the loss of information. We need to eradicate the bias. For sample, here recurring non-genuine transitions are added from 492. A total of observations should give an error of 3%. In this context instead of removing, adding a more number of observations. Hence, by utilizing this technique, we can achieve a high-accuracy model.

3. EXISTING SYSTEM

The existing systems for credit card fraud detection typically rely on rule-based systems or statistical models. Rule-based systems use predefined rules to detect fraud, such as setting a threshold for the maximum transaction amount or the number of transactions per day. Statistical models, on the other hand, use machine learning techniques to analyze historical transaction data and identify patterns of fraudulent behavior. Both of these approaches have limitations. Rule-based systems can be inflexible and fail to detect new or evolving fraud patterns. Statistical models, on the other hand, may require a large amount of labeled training data and can struggle to generalize to new data. To overcome these limitations, some existing systems have begun to incorporate deep learning techniques, which can automatically learn representations of data and identify complex patterns. Deep learning models have shown promising results in credit card fraud detection, achieving high levels of accuracy and reducing false positives. However, despite the advances in credit card fraud detection systems, fraudsters continue to develop new and sophisticated tactics to evade detection. As such, there is a need for ongoing research and development of more advanced and adaptive fraud detection systems to stay ahead of these evolving threats.

4. PROPOSED SYSTEM The proposed system for credit card fraud detection using RF, LR, DT, and SVM aims to evelop a more accurate and robust system by combining the strengths of multiple classifiers. The system involves collecting and preprocessing a large dataset of credit card transactions, followed by feature engineering to extract relevant features that can be used to train machine learning classifiers. Next, multiple classifiers will be trained, including Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM). These classifiers have been chosen based on their proven effectiveness in detecting fraud and their ability to work well together. The classifiers will be trained on the preprocessed data using a portion of the dataset and validated using another portion to ensure that the model does not overfit. The best performing classifiers will be selected based on their accuracy, precision, and recall. Overall, the proposed system offers the potential to significantly improve credit card fraud detection by leveraging the strengths of multiple classifiers, resulting in a more accurate and robust system that can adapt to new and evolving fraud patterns. Moreover, the system will be designed to be adaptable to new and evolving fraud patterns. As fraudsters continue to develop new tactics to evade detection, the system will need to be updated to ensure that it remains effective. This may involve updating the features used to train the classifiers or using new machine learning techniques to detect new patterns. In summary, the proposed system for credit card fraud detection using RF, LR, DT, and SVM has the potential to significantly improve the effectiveness of fraud detection systems. By leveraging the strengths of multiple classifiers, the system can achieve higher accuracy and robustness, while remaining adaptable to new and evolving fraud patterns. However, it is important to also consider the ethical implications of these systems and ensure that they are designed with fairness and transparency in mind.

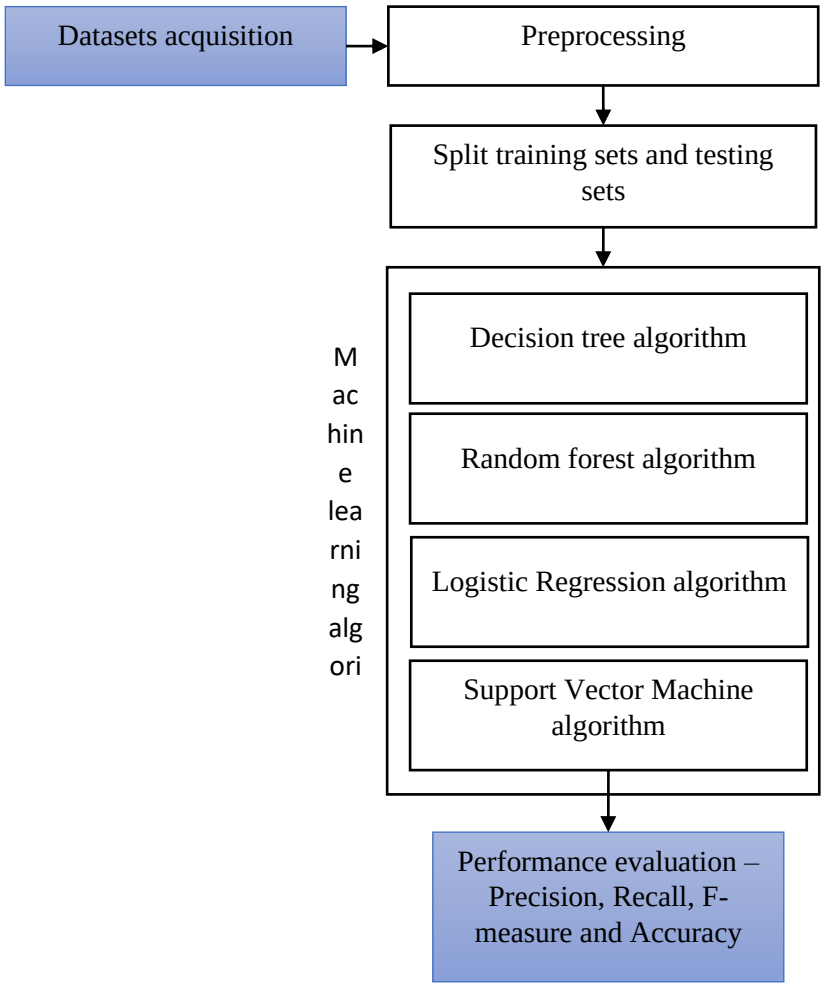

Fig 2: Proposed architecture

5. EXPERIMENTAL RESULTS

In this study, we can analyze classifiers in credit card fraud detection process that are performance in this project. Table 1 shows the fake news datasets attributes

ALGORITHM	F1 SCORE
Decision tree algorithm	60%
Random forest	70%
Logistic Regression algorithm	75%
Support Vector Machine	90%

algorithm	

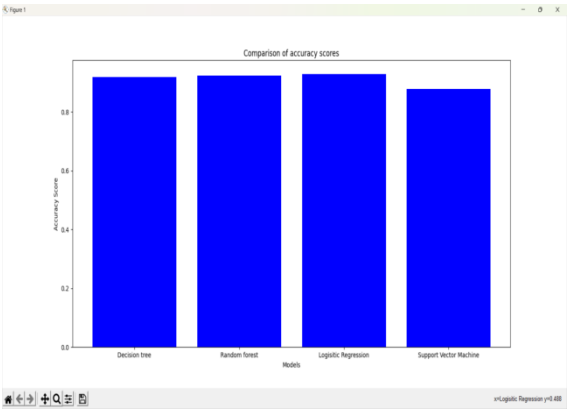

Fig 3: Performance chart

From this performance chart, proposed system provides improved F1 score than the existing machine learning classifiers.

6. CONCLUSION

In this work, we investigated the topic of detecting false news articles, creators, and subjects. Based on the news-augmented heterogeneous social network, explicit and latent traits may be retrieved from the textual information of news pieces, producers, and subjects. Furthermore, based on the relationships between news pieces, creators, and news subjects, a deep diffusive network model has been developed to include network structure information into model learning. Deep learning model improves accuracy ra In conclusion, credit card fraud is a serious problem that affects both consumers and businesses. The use of machine learning algorithms such as Random Forest, Decision Trees, Logistic Regression, and Linear SVM can help to detect fraudulent transactions and reduce the impact of credit card fraud. By combining these algorithms, we can create a more robust and accurate fraud detection system that is able to handle both linearly separable and non-linearly separable data. Additionally, the proposed system has the advantage of being able to handle imbalanced datasets and providing a high degree of accuracy in identifying fraudulent transactions. Overall, the proposed system has the potential to improve the efficiency and effectiveness of credit card fraud detection, thereby reducing the financial losses associated with fraudulent transactions. From the results, Logistic regression algorithm provide improved accuracy in classification for credit card datasets.

REFERENCES

[1] Sahu, Aanchal, G. M. Harshvardhan, and Mahendra Kumar Gourisaria. "A dual approach for credit card fraud detection using neural network and data mining techniques." In 2020 IEEE 17th India council international conference (INDICON), pp. 1-7. IEEE, 2020.

[2] Panda, Agyan, Bharath Yadlapalli, and Zhi Zhou. "Credit card fraud detection through machine learning algorithm." Big Data and Computing Visions 1, no. 3 (2021): 140-145.

[3] Aziz, Amir, and Hamid Ghous. "Fraudulent Transactions Detection in Credit Card by using Data Mining Methods: A Review." INTERNATIONAL JOURNAL OF SCIENTIFIC PROGRESS AND RESEARCH (IJSPR) 79, no. 179 (2021).

[4] Shah, Ankit, and Akash Mehta. "Comparative Study of Machine Learning Based Classification Techniques for Credit Card Fraud Detection." In 2021 International Conference on Data Analytics for Business and Industry (ICDABI), pp. 53-59. IEEE, 2021.

[5] Mienye, Ibomoiye Domor, and Nobert Jere. "Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions." IEEE Access (2024).

[6] Singh, Gurpreet, Divyanshi Kaushik, Hritik Handa, Gagandeep Kaur, Sunil Kumar Chawla, and A. Ahmed. "BioPay: a secure payment gateway through biometrics." Journal of Cybersecurity and Information Management 7, no. 2 (2021): 65-76.

[7] Tiwari, Pooja, Simran Mehta, Nishtha Sakhuja, Jitendra Kumar, and Ashutosh Kumar Singh. "Credit card fraud detection using machine learning: a study." arXiv preprint arXiv:2108.10005 (2021).

[8] Kumar, Sheo, Vinit Kumar Gunjan, Mohd Dilshad Ansari, and Rashmi Pathak. "Credit Card Fraud Detection Using Support Vector Machine." In Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021, pp. 27-37. Springer Singapore, 2022.

[9] Vinaya, D. S., Satish B. Basapur, Vanishree Abhay, and Neetha Natesh. "Credit Card Fraud Detection Systems (CCFDS) using Machine Learning (Apache Spark)." (2020).

[10] Lucas, Yvan, and Johannes Jurgovsky. "Credit card fraud detection using machine learning: A survey." arXiv preprint arXiv:2010.06479 (2020).