Credit Card Fraud Detection Using Machine Learning Algorithms

Varun Chaudhari
Department of Computer Engineering
Bharat College of Engineering
Badlapur, Thane, Maharashtra, India.
chaudharivarun2@gmail.com

Sanskar Gawade
Department of Computer Engineering
Bharat of College of Engineering
Badlapur, Thane, Maharashtra, India.
sanskargawade85@gmail.com

Kanay Shirke
Department of Computer Engineering
Bharat College of Engineering
Badlapur, Thane, Maharastra, India.
kanayshirke@gmail.com

Prof.Tushar Ubale
Department Of Computer Engineering
Bharat College of Engineering
Badlapur, Thane, Maharashtra, India.
tubale@bharatedu.co.in

***Abstract*—This paper introduces a novel approach to credit card fraud detection, leveraging a machine learning-based engine empowered by genetic algorithms for feature selection. Evaluating performance on a dataset derived from cardholders, our method surpasses existing systems, employing Decision Trees, Random Forest, and classifiers to achieve superior results in fraud detection. Expandability and Confidentiality issues that arise when incorporating these methods into actual financial security structures the research finds the best fraud detection algorithm by comparing random forests based on accuracy all things considered this study advances technologies for detecting credit card fraud and deals with the increasing difficulties brought by on by internet transactions.**

Keywords—Machine learning, Decision tree, Random Forest, Logistic Regression.

I. INTRODUCTION

Credit The exponential growth of the Internet over the last decade has led to a surge in online services such as e-commerce and tap-and-pay systems, consequently increasing the risk of credit card fraud. Despite the implementation of security measures like data encryption and tokenization, fraudsters continue to exploit vulnerabilities in credit card transactions. With credit card fraud posing significant challenges to financial institutions and individuals worldwide, there is an urgent need for more effective detection methods. Traditional rule-based systems and manual assessments are proving inadequate in addressing the evolving nature of fraudulent activities. Machine learning (ML) algorithms offer promising solutions by leveraging data-driven approaches for fraud detection. However, challenges persist due to the confidential nature of credit card transaction datasets and the constantly changing patterns of fraudulent behavior. This research explores the application of supervised ML algorithms such as Decision Trees, Random Forests, and Logistic Regression for credit card fraud detection. By utilizing large datasets and the aim is to develop models capable of accurately detecting fraudulent transactions and minimizing financial losses.

II. LITERATURE SURVEY

A. *Sampling the imbalance dataset*

One of the difficulties faced by researchers was balancing an unbalanced dataset. The issue arose from the inability of the current methods to adjust to the dataset's imbalanced class composition. However there are several approaches to identifying an imbalance, and they can be further differentiated according to the algorithms applied and the volume of data handled (Bagga et al., 2020). Unbalanced data is first pre-processed at the level of data detection to eliminate noise from the data and apply the appropriate method. The following categories can be used to group the data-leveling techniques:

- Cost-based: samples the data to lower the system's overall computing complexity; samples the data to prevent over-fitting
- Distance-based: take a sample of the data and replicate the various distances that are related to each class.

In contrast, classifiers typically identify the minority class when attempting to differentiate a dataset depending on the algorithms employed (Bagga et al., 2020).

- Leveling out the unbalanced dataset
- Reducing computing complexity with machine learning-based classifiers

The author's research study (Tingfei, H., Guangquan, C., and Kuihua, et al., 2020) found that an imbalance in the dataset caused the system model's performance efficiency to drop. He suggested using machine learning techniques like SVM, random forest, and KNN for this reason. Although the model's overall efficiency increased, the problem of data imbalance remained unresolved. In a related study the authors (Taha, A. A., and Malebary, S. J., 2020) as well as (Arya et al., 2020); also made contributions to the same field and encountered issues with an imbalance in the dataset.

B. *Algorithmic techniques*

The authors (Sadgali et al., 2020) suggested utilizing machine learning to detect credit card fraud. The pre-processed data was used before the algorithms were used. After obtaining the dataset from the relevant source, the information was cleaned, and filtered to remove duplicates, and any irrelevant information was eliminated. As a result, the quantity of data required to run machine learning algorithms was limited. The following machine learning algorithms, including KNN, SVM, and logistic regression, were used by the authors in their work. Initially, the imbalanced data concerns were addressed, and a strong data visualization display was seen. Apart from the diverse range of machine learning algorithms employed, the writer additionally suggested CNN as the neural network, encompassing 21 layers of undiscovered neurons. The addition of hidden layers improved the system model's overall accuracy, resulting in an accuracy of 77.96 percent.

In a related study, the authors (Saheed YK et al., 2020) put forth a principle that relied on a decision tree and a functional set of probability rules. A voting system was used to create decision trees, and the class with the most votes was chosen as the optimal classifier. The author suggested implementing three machine learning-based classifiers and one stacking-based method in addition to the voting classifier. It was suggested that stacking be used as the hybrid classifier to produce a maximum accuracy of 81.35 percent, which resulted in the of undiscovered neurons. The addition of hidden layers improved the system model's overall accuracy, resulting in an accuracy of 77.96 percent.

In a related study, the authors (Saheed YK et al., 2020) put forth a principle that relied on a decision tree and a functional set of probability rules. A voting system was used to create decision trees, and the class with the most votes was chosen as the optimal classifier. The author suggested implementing three machine learning-based classifiers and one stacking-based method in addition to the voting classifier. It was suggested that stacking be used as the hybrid classifier to produce a maximum accuracy of 82.35 percent, which resulted in the creation of an optimized model. The overall implementation of credit card fraud detection was accomplished, and a filtered dataset was obtained.

C. Feature Selection Strategies

Using the concepts of feature selection, authors (Seera M et al., 2021) conducted an implementation of five classifiers using random forest, SVM, KNN, decision trees, and Naïve Bayes. Redundant information was initially eliminated for a better understanding of the algorithms. Using feature selection, the time complexity, speed, and performance were elevated, and the purpose of detecting credit card fraud was eventually achieved.

The authors (Khatri S., Arora A., and Agrawal AP, 2020) proposed the implementation of SVM and logistic regression alone to predict and detect the occurrence of credit card fraud. The dataset used for the same could further classify the obtained training and test files as fraudulent and non-fraudulent. The authors added the implementation of SMOTE along with the feature selection method. SMOTE was a means of correcting the dataset's imbalance. Machine learning classifiers were fed an equal distribution of fraudulent and non-fraudulent examples, allowing for the measurement of the classifiers' performance and the creation of an improved model. Apart from implementing SMOTE, the authors additionally employed a random forest technique to finalize the feature selection procedure. This involved selecting just pertinent characteristics and discarding the remaining ones to preserve overall efficiency.

D. Missing Value Strategies

There are situations where some datasets include missing values, which could result in a reduction in the system model's overall performance. This dataset's missing values could potentially replace NULL values, which would cause a drop in data by leaving gaps in the columns. The total accuracy and precision would be much reduced if these data were used for evaluation or training. Therefore, methods for overcoming them are also required.

To identify credit card fraud, the authors (Varmedja D et al., 2019) suggested using SVM, logistic regression, and KNN. They used bar graphs to illustrate the data after obtaining the dataset from the Kaggle repository for this purpose. First, an imbalance in the data was noticed, wherein numerous instances that weren't fake were observed. The authors used SMOTE techniques to get around this. Subsequently, it was noted that the gathered dataset had NULL and missing values. The day came when these values had to be removed and eliminated. Afterwards, the author's data-preprocessing methods took care of this. Pre-processing methods included transformation and cleansing of the data. Using machine learning-based classifiers on a filtered dataset, the system obtained a 75.63 percent accuracy rate through logistic regression.

E. Research

TABLE 1. Literature Review Analyses of Various Papers

Authors	*Classifiers Used*	*Accuracy*	*Advantages*	*Drawbacks*
(Robles-Velasco A et.al, 2020)	SVM, KNN, and Logistic Regression	71.28 percent	Feature extraction and selection methods were performed to filter redundant data	Precision values were less in comparison to the implementation of other algorithms
(Maniraj SP et, al. 2019)	Decision trees, random forests, Naïve Bayes, and KNN	76.32 percent	The voting mechanism of helped to increase the accuracy of the overall system	The acquired dataset had a restricted number of cases
(Lingjun H et.al, 2020)	CNN and Naïve Bayes	79.56 percent	Increased efficiency due to hidden layers	Significant challenges of data imbalance occurred

III. METHODOLOGY

A. Logistic Regression

One supervised machine learning approach that is commonly used for binary classification is called logistic regression. Predicting the likelihood that an instance falls into a specific class—such as spam or not, sickness or not—is its main objective. Logistic regression is more concerned with categorical outcomes than linear regression, which forecasts continuous values. The sigmoid function, sometimes referred to as the logistic function, is used in logistic regression to translate input variables into probabilities ranging from 0 to 1.

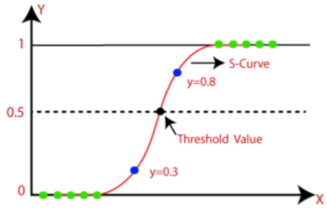

Fig 1. Logistic Regression

B. *Random Forest*

Random Forest uses a decision tree algorithm in a random manner or in a random way. It is ensemble learning used by both classification and regression. It combines the output of multiple decision trees to get a single Result. It is easy to use and flexible to handle.

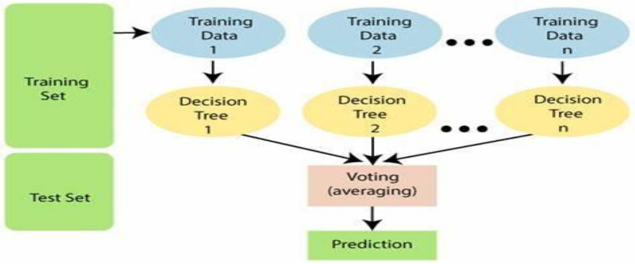

Fig 2. Representation Of Random Forest

C. *Decision Tree*

A decision tree is one of the types of supervised algorithms used for both type classification and regression duties. It builds a flowchart-like tree shape in which every internal node denotes a check on an attribute, every branch represents an outcome of the take-a-look-at, and each leaf node holds a category label. It's constructed by recursively splitting the education data into subsets primarily based on the values of the attributes until a preventing criterion is met, which includes the maximum intensity of the tree or the minimum range of samples required to split a node.

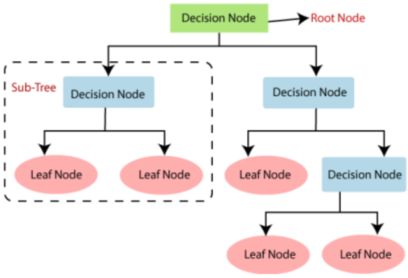

Fig 3. Representation of Decision Tree

IV. SYSTEM DESIGN AND DESIGN SPECCIFICATION

A. *System Design*

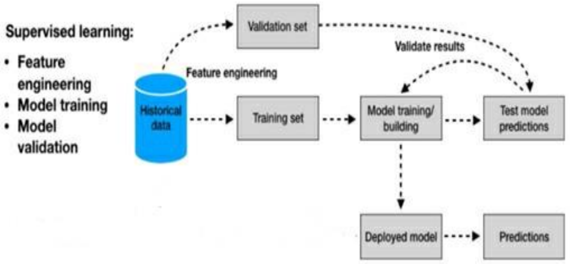

Fig 4. System Design of CCFD using Machine Learning

The architecture of the system design begins with the collection of the dataset from repositories that contain fraudulent transactions gathered from banks and various payment sectors. Such a monetary transaction-related database can further be categorized into the following features:

- Features related to accounts: such features include the account number of the user, information on his card limit, account opening date, CVV, etc.
- Features related to monetary transactions: such features include information related to the amount being transferred, the transaction time, details of the recipient, etc.

After the collection of the dataset, feature engineering techniques are applied, and the data is further fed for training using machine learning classifiers. After the completion of this stage, the dataset is validated and put into the testing phase for results. In contrast to historical data, when a new data upload occurs in the database, the same feature engineering process takes place on the dataset, and training occurs. Once the model is deployed, it is finally used for prediction.

During its actual execution, a transaction is made by the customer at a specific time. On the back end of the system model, all the occurring transactions are stored in the database and form a set of historical data.

To detect the occurrence of fraud, it is necessary to assume the legitimacy of all monetary transactions. This is done so that the system model can further classify the data as legitimate or fraudulent. The system design for machine learning-based implementation can further be categorized into three stages:

1. Building a system model from scratch using historical data.
2. Predicting the system model using machine learning-based classifiers.
3. Evaluating the model using performance measures.

In simple terms, credit card fraud detection is like having a smart system that helps protect your money. It watches out for any tricky or dishonest activities related to your credit cards. It does this by looking at your transactions and checking if they seem normal or if something looks fishy. If it sees anything suspicious, like someone using your card without permission, it alerts you or the bank so they can stop the bad guys from taking your money.

B. *Design Specifications*

The design specifications are the prerequisites needed to run any machine learning algorithm it consist mostly of the requirements that serve as the cornerstone of any system model the majority of the study's design specifications may seem similar requiring modifications solely for the unique way that Python modules are used the suggested thesis design specifications are as follows. Following are the design specifications of the proposed thesis:

1. Languages used to implement: Python is chosen as the base language, and the implementation occurs using Python features.
2. Libraries used to implement: Python libraries such as Pandas are used to load the respective data of CCFD.
3. Hardware used to implement: the proposed study demanded hard disk storage with 12 GB of RAM.

V. DATA UNDERSTANDING AND PREPARATION

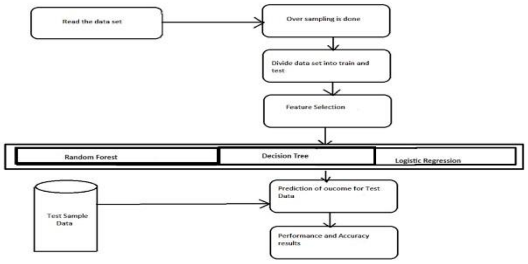

Fig 5. Workflow of the Proposed Study

The research process initiates with acquiring data from the Kaggle repository. Following this, the dataset is subjected to a preprocessing phase, where techniques like SMOTE are employed to ensure balance. Subsequently, the data undergoes extraction and selection to eliminate redundancies. Visualization methods are then applied to gain insights into the data. Next, the dataset is split using four distinct machine-learning algorithms alongside a stacking algorithm. The resulting test data is utilized for CCFD prediction. Finally, the efficiency of the classifiers is evaluated using various performance metrics.

A. Dataset Used

The dataset sourced from Kaggle encompasses transactions from January 2019 to December 2020, featuring monetary exchanges from 1000 customers and involving a diverse array of 800 merchants. This dataset, created via the Sparkov data generator, amalgamates legitimate and fraudulent cases, facilitating a comprehensive analysis. The dataset is comprised of 46 columns delineating various attributes, including transaction dates, credit card details, merchant information, and demographic data such as occupation, gender, street, and zip code.

B. Data Pre-Processing

The dataset retrieved from the Kaggle repository initially comprises unprocessed, unbalanced, and raw data, prolonging training periods and complicating system computations, necessitating the eradication of data redundancy. Subsequently, through data pre-processing, redundant entries are removed, NULL values are normalized, relevant columns and their attributes are selected, and the dataset is balanced using SMOTE, collectively enhancing the efficiency of the model's computations.

C. Balancing the Data

To address the inherent inequality in the dataset sourced from the repository, it's essential to balance it before implementing appropriate methodologies. Utilizing SMOTE, or Synthetic Minority Oversampling Technique, offers a straightforward approach to replicate instances of the minority class, effectively resolving data imbalance issues. This oversampling strategy enables the generated samples to closely align with the model's expectations, thus fostering a balanced dataset conducive to accurate analysis. Through SMOTE's replication technique, the synthesized samples can be compared with the minority samples, enriching the dataset's diversity and enhancing its representativeness.

Classifiers	***Accuracy***
Logistic Regression	50%
Decision Tree	85%
Random Forest	96%

D. Data Visualization

Data visualization is a process in which the obtained data from the repository can be put to visualize using bar graphs and plots. Such graphs help to analyze historical data gained from past events and further make decisions so that respective predictions can be made.

VI. RESULT

All of the three classifier the accuracy of random forest Algorithm is higher than the other classifier so we than selected random forest algorithm has our primary Algorithm.

TABLE 2. Classifier and Accuracy

VII. CONCLUSION

The exponential growth of online transactions has necessitated the implementation of sophisticated fraud detection mechanisms to combat increasingly intricate fraudulent activities. Traditional rule-based systems and manual assessments have proven inadequate in addressing the difficult nature of credit card fraud, leading to a growing reliance on machine learning algorithms for automated detection. While various supervised learning algorithms such as Decision Trees, Random Forests, and Logistic Regression have been explored, challenges persist due to the confidential nature of transaction datasets and the skewed distribution of fraudulent instances. Nonetheless, advancements in feature selection techniques, ensemble methods, and model evaluation strategies offer promising avenues for enhancing the accuracy and efficiency of fraud detection systems. Moving forward, continued research and collaboration between academia, industry, and regulatory bodies are essential for developing robust, adaptive, and scalable solutions capable of effectively combating evolving forms of credit card fraud while minimizing false positives and ensuring user privacy and security.

REFERENCE

[1] Varun Kumar KS, http://www.ijert.org: Fraud Detection using Machine Learning in e-Commerce (IJACSA)", International Journal of Advanced Computer Science and Applications, vol. 10, no. 9, 2019.

[2] Heat Naik and Prashant Kanika, "Credit card Fraud Detection based on Machine Learning Algorithms", International Journal of Computer Applications (0975 - 8887), vol. 182, no. 44, March 2019.

[3] Radula Kora age, "Faculty of Information Technology" in Data Mining Techniques for Credit Card Fraud Detection, University of Moratuwa.(2020).

[4] Radula Kora age, "Faculty of Information Technology" in Data Mining Techniques for Credit Card Fraud Detection, University of Moratuwa.(2021).

[5] Unam Sam, Godfrey Moses, Taiwo OlajideCredit Card Fraud Detection Using Machine Learning Algorithms(2023).

[6] Emmanuel Ileberi1*, Yanxia Sun1 and Zenghui Wang2 A machine learning-based credit card fraud detection using the GA algorithm for feature selection (2022).