Credit Card Fraud Detection Using Machine Learning

Shital.S. Pawar[1], Vaibhavi Dhapate[2], Aniket Lahane[3], Prof. Sonali K. shewale[4]
[1,2,3] *P.E.S. College of Engineering Chh.SambhajiNagar, Maharashtra, India*
[4]*Guide, Department of Computer Science & Engineering D-Batu (Dr. Babasaheb Ambedkar Technological University), Raigad, Lonere.*

***Abstract*—Credit card fraud poses a significant concern for both financial institutions and their customers. To combat this issue, researchers have explored numerous techniques, including machine learning and deep learning, to develop effective credit card fraud detection systems.[6] credit card fraud detection system using machine learning, thereby addressing the growing financial risks associated with increased transaction volumes.**

Utilizing a supervised learning approach on historical transaction data, the methodology encompasses data preprocessing, model selection.

Techniques to handle imbalanced datasets, such as SMOTE, are employed. The expected outcome is a high-accuracy model that effectively distinguishes fraudulent transactions, minimizing false positives and negatives, and improving overall fraud detection rates."

I. INTRODUCTION

The cardholder is then billed periodically for these charges. Modern systems like ATMs, store readers, banks, and online banking platforms can all read the information stored on the card. Each credit card has a unique card number, which is vital for its security.

Credit cards have undeniably streamlined digital transactions, making them more accessible than ever. However, this convenience comes with a significant drawback: criminal credit card transactions result in billions of dollars in losses annually. A 2017 PwC global economic crime survey highlighted the scale of this issue, indicating that roughly 48% of organizations had experienced economic crime.[5] As a result, it's crucial to develop robust solutions for credit card fraud. Furthermore, the rapid evolution of new technologies unfortunately provides novel avenues for criminals to perpetrate scams. These fraudulent activities inflict substantial financial damage not only on merchants and banks but also directly on individual cardholders.

This complex issue warrants significant attention from the machine learning and data science communities, as it presents a prime opportunity for automated solutions. This paper aims to provide a structured overview of the current state of credit card analytics, drawing insights from a wide range of scholarly articles, research papers, and industry reports.[3] In practical scenarios, automated tools rapidly process the massive incoming flow of payment requests, instantly deciding which transactions to approve. Machine learning algorithms then scrutinize all authorized transactions to flag any suspicious activity. These flagged instances are subsequently investigated by human experts who contact cardholders to verify the legitimacy of the transactions.

Fraud detection methodologies are constantly evolving to counter the dynamic tactics employed by criminals.

These fraudulent activities can typically be categorized as follows: [2]

- Credit Card Misuse
- Card Compromise
- Account Exploitation
- Digital Device Intrusion
- New Account Fraud
- Forged Cards
- Telecommunications Scams

Some of the prominent techniques currently employed for detecting such fraud include: [2]

- Artificial Neural Networks (ANNs)
- Genetic Algorithms
- Logistic Regression
- Decision Trees
- Fuzzy Logic
- Support Vector Machines (SVMs)
- Bayesian Networks

A credit card fraud occurs when someone uses another person's card for their own personal usage without the owner's knowledge.[4]

2.SYSTEM ARCHITECTURE

Fig. System Architecture

3.LITERATURE REVIEW

In its most basic form, fraud involves illicit or criminal trickery carried out to achieve financial or individual gain. It's a deliberate transgression of law, rule, or policy, specifically with the aim of acquiring unlawful monetary benefit. The field of anomaly or fraud detection has seen extensive research, much of which is publicly available.

In their thorough survey, Clifton Phua and colleagues emphasized that this field utilizes approaches like data mining applications, automated fraud detection, and adversarial detection. Furthermore, novel approaches like the hybrid data mining/complex network classification algorithm have proven effective for medium-sized online transactions.

This algorithm relies on a network reconstruction method that builds representations of how individual instances diverge from a reference group. Despite these advancements, fraud detection remains a significant challenge, and no single algorithm can flawlessly predict whether a transaction is fraudulent. An effective fraud detection system should accurately identify fraudulent activities, detect them swiftly, and crucially, avoid misclassifying legitimate transactions as fraudulent.

4.METHODOLOGY

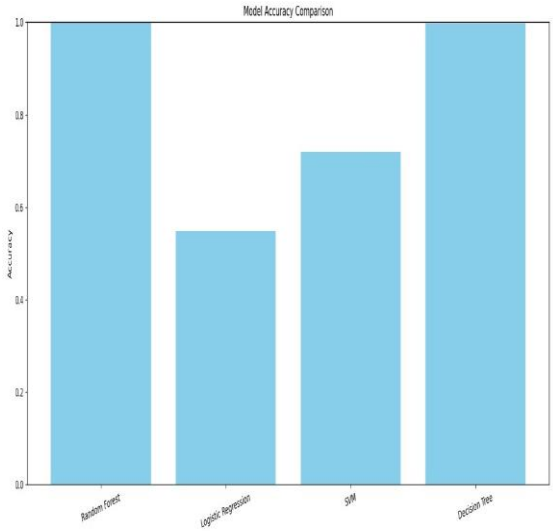
This paper introduces an approach that leverages cutting-edge machine learning algorithms to identify anomalous activities, often referred to as outliers. A preliminary architectural overview is depicted in the accompanying figure.

To begin, our dataset was obtained from Kaggle, a well-known source for analytical datasets. This dataset is structured with 11 columns.

The 'Category' column representing a genuine transaction category like on which we have done transaction like, food, fuel, shopping etc.

The column transaction type demonstrates the type of transaction as withdrawal, credited or debited. The columns for Date, Day, Month and Time in hours on which transaction is done successfully using credit card are there.

We generated various graphs to visually inspect the dataset for any inconsistencies and to aid in its overall comprehension.

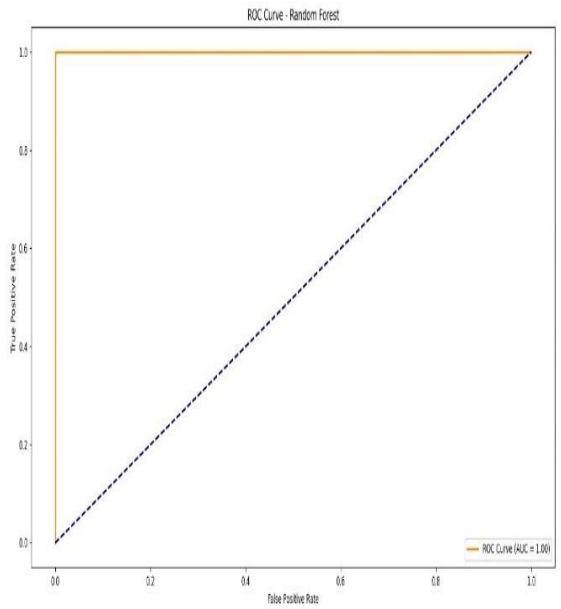


This graph shows the accuracy of different models: Logistic Regression, Decision Tree, Random Forest and SVM.

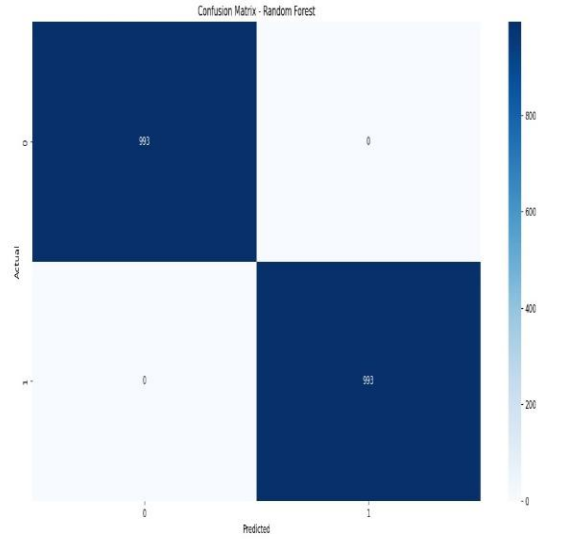
Here's an analysis of the accuracy for each model based on the graph:

- Random Forest: The accuracy is at 1.0 (or 100%).
- Logistic Regression: The accuracy is approximately 0.55 (or 55%).
- SVM: The accuracy is approximately 0.72 (or 72%).
- Decision Tree: The accuracy is at 1.0 (or 100%).

From the graph, both Random Forest and Decision Tree models achieved the highest accuracy (1.0), while Logistic Regression had the lowest accuracy among the four.



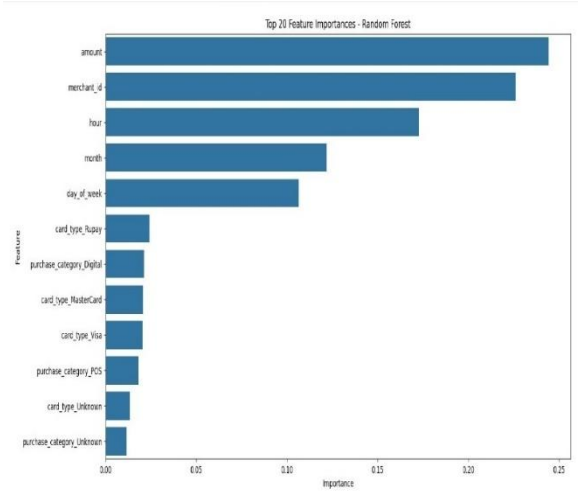
The graph shows an ROC Curve (Receiver Operating Characteristic Curve) for a Random Forest model, and it shows an Area Under the Curve (AUC) of 1.00.



The graph is a Confusion Matrix for a Random Forest model. This matrix is extremely valuable for understanding the performance of a classification model.

The confusion matrix has four quadrants: True Negatives (TN), False Positives (FP), False Negatives (FN), True Positives (TP)

A lesser false positive rate signifies a lower occurrence of incorrect alerts, enhancing the reliability of positive identifications.



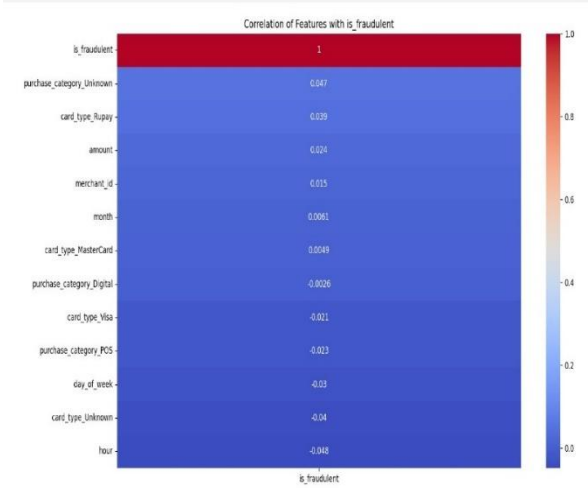
This graph shows the Top 20 Feature Importances for a Random Forest model.

It's a horizontal bar chart shows the

Y-axis: list different "Features" used by the model and

X-axis: represents the "Importance" score for each feature.

In this graph Based on the features listed (amount, merchant _id, hour, card _type, purchase_ category), it strongly suggests that the model is likely used for fraud detection or a similar transactional analysis task.



This graph is a correlation matrix showing the Correlation of Features with is fraudulent. The core analysis centers on mapping the associations between various data characteristics and the 'is_ fraudulent' outcome.

This suggests the graph is part of an analysis for a fraud detection model, where is fraudulent is likely a binary variable (e.g., 1 for fraudulent, 0 for non-fraudulent).

The Y-axis lists different features (e.g., purchase_ category_ Unknown, card_ type_Rupay, amount, merchant_ id, month, hour, etc.).

5.IMPLEMENTATION

The real-world application of this idea faces substantial hurdles. Banks' unwillingness to share data, driven by competitive market dynamics, legal constraints, and the need for user data protection, makes practical implementation particularly difficult. Our research consequently involved an examination of pertinent literature that employed analogous methodologies and reported corresponding results. As indicated in one of these referenced publications. For banking confidentiality reasons, only a summary of the results obtained is presented below.[7] After applying this technique, the level 1 list encompasses a few cases but with a high probability of being fraudsters.[7] we have done in python by importing python libraries like pandas, numpy, from sklearn model import train test split, Random Forest Classifier, Logistic Regression, SVC, Decision Tree Classifier, accuracy score, Standard Scaler, SMOTE, make pipeline, pickle.

1.Data Preprocessing:

In These Code Initially we loaded the dataset, and dropped unused columns. After That filled missing values with mode/median. The dataset's class imbalance is resolved through the application of SMOTE.

2.Model Selection:

In this section we separate the Dataset for training and testing as training dataset and testing dataset in the ratio of 0.8 and 0.2.

We trained Random Forest, Logistic Regression, SVM, and Decision Tree models on training dataset. Evaluated model's accuracy based on their predictions, and selected the one with highest accuracy as best model.

3.Prediction on selected accurate model:

This system saves the optimal model and its feature columns, then predicts fraud probability for new, user input transactions after aligning them with the training data.

Improving time efficiency and minimizing overhead charges could be achieved by introducing new fields into the query structure. Examples of these new query elements include the first five digits of phone numbers, email addresses, or passwords. These updated queries could be effectively applied to both the level 2 and level 3 lists.

6.RESULT

Our implementation outputs the number of false positives identified, which is subsequently validated against actual values to derive the accuracy and precision scores for the algorithms. For quicker preliminary testing, we processed 10% of the total dataset. Ultimately, the entire dataset is also run, and both sets of results are displayed.

The ensuing output presents these findings, along with a detailed classification report for each algorithm. Here, 'class 0' denotes a legitimate transaction, and 'class 1' signifies a fraudulent one. The accuracy of the detection is confirmed by comparing these results against the known class values, specifically checking for false positives.

7. CONCLUSION

Credit card fraud fundamentally constitutes a deceptive criminal offense. This article explored common fraud methods and their detection techniques, while also reviewing current research. Significantly, it detailed how machine learning can boost fraud detection results, providing insights into the algorithm, pseudocode, explanation, and experimental outcomes. The other way to handling imbalance dataset is to use one-class classifiers like one-class SV.[1]

REFERNCE

[1] Credit Card Fraud Detection using Machine Learning Algorithms - ScienceDirect
[2] (PDF) Credit Card Fraud Detection using Machine Learning and Data Science
[3] (PDF) Credit Card Analytics: A Review of Fraud Detection and Risk Assessment Techniques
[4] (PDF) CREDIT CARD FRAUD DETECTION SYSTEM
[5] CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING
[6] JETIR2307462.pdf
[7] Credit Card Fraud Detection using Machine Learning and Data Science