IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

UPI Fraud Detection Using Machine Learning

[1]Sumit M. Kulkarni, [2] Prof. Jyoti Gavhane
[1]Student [2]Professor
[12]MIT School of Computing, MIT-ADT University, Pune, India

Abstract: This project presents a full-stack web application designed for the detection of UPI (Unified Payments Interface) fraud using advanced analytical techniques. The system leverages historical transaction data to identify patterns and anomalies indicative of fraudulent activities. The application is equipped with a user-friendly interface, allowing users to input transaction details and receive real-time fraud assessments. By employing a robust backend framework and a responsive frontend, the application ensures seamless interaction and effective data management. Additionally, the project emphasizes data security and user privacy, adhering to best practices in web development. The outcome is a comprehensive solution that enhances the integrity of digital transactions and provides users with confidence in their financial activities.

Index Terms - UPI Fraud Detection, Data Security, Unified Payments Interface, Transaction Data Analysis.

I. INTRODUCTION

The Unified Payments Interface (UPI) has become a cornerstone of digital payments in India, simplify financial transaction and contributing significantly to the shift toward a cashless economy. However, the rapid adoption of UPI has also led to a concerning rise in fraudulent activities, posing serious risk to user security and the trustworthiness of digital financial platforms. Fraudulent practices sucj as phishing attacks, unauthorized transaction, and malware-based exploits have made it imperative to adopt advances solutions for fraud detection and prevention.

The rise in UPI-related fraud underscores the necessity for real-time, accurate, and efficient fraud detection systems. Traditional fraud detection methods are often inadequate, as they are prone to errors and delays, which can result in significant financial losses for usets and service providers. Machine Learning (ML) has emerged as a powerful tool to address this challenge by enabling systems to detect fraudlent patterns with high precision, ensuring the safety and reliability of digital transactions.

According to the Reserve Band of India, UPI has recorded billions of transactions, with the volume and value of transactions growing exponentially. However, this surge in activity has also amplified the complexity and scale of fraud detection. Fraud and occure in various forms, including socual engineering attacks, account takeovers, and payment reversals, each exploiting different vulnerabilities in the system. These activities not only

This research is motivated by the increasing prevalence of UPI fraud and the need for robust security mechanisms. It aims to leverage machine learning to enhance fraud detection, focusing on algorithms such as Random Forest, Logistic Regression, Decision Tree, and Support Vector Machine(SVM). These algorithms are designed to classify transactions as legitimate or fraudulent based on transaction patterns and behavioral indicators. Additionally, the study emphasizes the importance of model transparency and interpretability in financial applications, enabling stakeholders to understand the factors contributing to fraud detection and ensuring accountability in decision-making.

The goal of this research include developing a machine learning-based UPI fraud detection model, improving fraud detection accuracy while minimizing false positives, and assessing the performance of different ML alorithms. Furthermore, the study addresses challenges such as imbalances dataset in fraud detection, identifies critical features influencing fraudulent activities, and purposes and end-to-end fraud detection system. By incorporating explainability techniques such as SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations), the study enhances the transparency and reliability of the proposed models, ensuring their practical applicability in securing digital transactions.

In summary, this research seeks to design a reliable and explainable machine learning model for UPI fraud detection, address challenges associated with imbalanced dataset, and develop an effective fraud prevention system. By doing so, it aims to fortify the UPI ecosystem, safeguard users against fraud, and maintain the integrity of India's digital payment infrastructure.

II. LITERATURE REVIEW

A. UPI Fraud Detection Using Machine Learning

Patil et al. introduced a fraud detection framework that combines behavioural analytics and Support Vector Machines (SVMs). Their approach analysed transaction data for anomalies, focusing on features such as transaction timing, geographical mismatches, and transaction velocity. This method effectively distinguished between legitimate and fraudulent transactions. However, the study highlighted interpretability challenges, particularly with non-linear kernels, which limited transparency in decision-making processes. Despite these challenges, SVMs achieved high precision in detecting anomalies.

B. UPI Fraud Detection Using Machine Learning

Shabreshwari et al. explored multiple machine learning algorithms, including Random Forest (RF), XGBoost, LightGBM, and Logistic Regression, to detect fraudulent UPI transactions. They emphasized feature engineering, identifying relevant attributes such as transaction amounts, timestamps, and user device information. Their ensemble-based models demonstrated high precision and recall, outperforming simpler classification models. By integrating real-time alert systems, their approach reduced false positives while ensuring the system remained responsive to new fraud patterns.

C. Fraud Detection in UPI Transactions Using ML

A study by Kavitha et al. employed Hidden Markov Models (HMMs) for UPI fraud detection. The HMM framework analyzed user-specific transaction patterns to establish behavioral baselines, flagging deviations as potential fraud. This approach proved particularly effective for personalized fraud detection but required computational resources and was sensitive to parameter selection. Nevertheless, the model demonstrated the adaptability of probabilistic methods to detect evolving fraud patterns.

D. UPI Fraud Detection Using Machine Learning

Imbalanced datasets, where fraudulent transactions constitute a small proportion of the total, are a core challenge in fraud detection. Techniques such as oversampling, undersampling, and synthetic data generation have been explored in prior studies. Shabreshwari et al. utilized these techniques to balance their datasets, which significantly improved model performance metrics such as recall and F1-score. These efforts highlight the importance of data preprocessing to enable machine learning models to handle rare event detection.

E. UPI Fraud Detection Using Machine Learning

Comparative analyses have provided insights into the strengths and weaknesses of various models for fraud detection. Shabreshwari et al. compared Decision Trees, Naive Bayes, Logistic Regression, and ensemble models, finding that ensemble methods such as Random Forest and XGBoost offered the best balance of accuracy and precision for fraud detection in UPI transactions. Logistic Regression with L1 regularization proved particularly useful for feature selection, ensuring simpler models with strong interpretability.

III. METHODOLOGIES:

1. Data Consideration:

The dataset for UPI fraud detection comprises features such as transaction amount, timestamp, user location, device ID, and user ID, with labels distinguishing between legitimate and fraudulent transactions. A well-structured dataset forms the foundation for building a reliable fraud detection model.

2. Data Preprocessing:

Preprocessing is a critical step to ensure the dataset is clean, consistent, and ready for analysis. It involves:
Handling Missing Values and Outliers: Missing values are addressed, and outliers are removed to maintain data integrity.
Data Scaling: Standardization is applied to normalize feature values, particularly important for algorithms like SVM that are sensitive to feature scaling.

3. Train-Test Split and Cross-Validation:

The dataset is divided into training and testing sets to evaluate the model's performance on unseen data.

K-Fold Cross-Validation: Ensures robustness by partitioning the dataset into multiple subsets, training and testing the model iteratively on these folds to prevent overfitting.

4. Algorithm Implementation:

- **Random Forest**: Implemented to handle non-linearity, leveraging ensemble learning for robust fraud detection.
- **Logistic Regression**: Applied for probability-based classification, helping understand transaction likelihoods.
- **Decision Tree**: Used for simple, interpretable decision-making, suitable for analyzing individual feature effects.
- **SVM**: Deployed for complex, high-dimensional data classification, maximizing separation between classes.

5. Evaluation Metrics:

- **Accuracy**: Measures the overall correctness of fraud detection.
- **Precision**: Indicates the model's accuracy in identifying true fraud cases.
- **Recall**: Measures sensitivity to actual fraud cases.
- **F1 Score**: Balances precision and recall for comprehensive performance analysis.
- **ROC-AUC**: Assesses the model's effectiveness in distinguishing between classes.

IV. ARCHITECTURE:

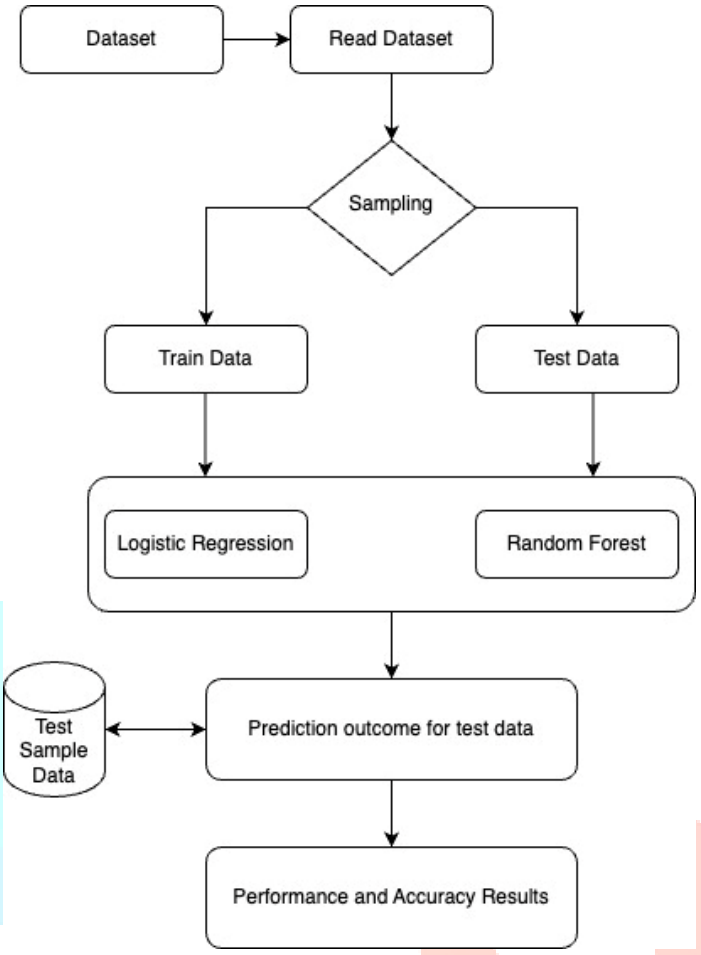

Fig. 1. Flowchart of UPI Fraud Detection

1. **Dataset**
   The process begins with the dataset, which provides the necessary input data for analysis. In applications like texture analysis or fraud detection, this dataset contains various features or attributes that are used for classification tasks.
2. **Read Dataset**
   The dataset is loaded into the system to ensure it is ready for processing and further analysis.
3. **Sampling**
   The dataset is divided into two parts:

   Training Data: Used for training the machine learning models.

   Test Data: Used for evaluating the performance of the trained models.
4. **Model Training**
   Two machine learning models are implemented simultaneously:

   Logistic Regression: A statistical model used for classification. It is effective for linearly separable data and provides easy interpretability.

   Random Forest: An ensemble learning method that utilizes multiple decision trees to improve accuracy and handle non-linear relationships in the data.
5. **Test Sample Data**
   Once the models are trained, unseen test sample data is introduced to evaluate their predictive capabilities.
6. **Prediction Outcome for Test Data**
   The trained models generate predictions for the test data. In fraud detection, for example, this step would involve distinguishing between legitimate and fraudulent transactions.
7. **Performance and Accuracy Results**

The models' performance is assessed using various metrics, including:

Accuracy: Measures the overall correctness of the predictions.

Precision: Focuses on the ratio of true positives among all predicted positives.

Recall: Evaluates the ability to correctly identify all relevant instances.

F1 Score: Combines precision and recall to provide a balanced measure of performance.

V. OUTCOME:

The evaluation of each model in detecting fraudulent transactions yielded the following outcomes:

Random Forest

Random Forest demonstrated high accuracy and precision. Its ensemble learning approach, which combines multiple decision trees, reduces the risk of overfitting, making it highly effective at generalizing across different types of fraud patterns. This allows for a more reliable detection of fraudulent transactions, even with varied data. The model's ability to handle complex, non-linear relationships in the data contributed to its strong performance.

Logistic Regression

Logistic Regression excelled in terms of interpretability. It provides probabilistic outputs that indicate the likelihood of a transaction being fraudulent. This makes it particularly useful for scenarios where fine-tuning the sensitivity of fraud detection is important. By adjusting the decision threshold, users can control the trade-off between identifying fraud cases and minimizing false positives, offering flexibility in real-time fraud monitoring.

Decision Tree

The Decision Tree model is valued for its simplicity and transparency. It generates clear decision paths, making it easy to understand how the model arrived at its classification. This model is especially useful for identifying the specific features or factors that contribute most to classifying a transaction as fraudulent or legitimate. However, Decision Trees may struggle with handling complex data patterns compared to other algorithms like Random Forest.

Support Vector Machine (SVM)

SVM showed clear classification boundaries between fraudulent and legitimate transactions. It works well with high-dimensional data and can efficiently classify complex patterns. However, SVM may require significant computational resources, which can make it less practical for real-time applications, especially when processing large datasets or in environments where speed is critical.

Comparison and Conclusion

Comparing these models suggests that Random Forest is the preferred choice for detecting fraud. It strikes a good balance between accuracy, precision, and interpretability, making it effective and versatile in real-world applications. However, Logistic Regression also offers value, particularly in settings where custom thresholds for fraud detection need to be adjusted based on the probability scores. While Decision Tree and SVM provide useful features, their limitations (simplicity and computational intensity, respectively) make them less optimal for comprehensive fraud detection when compared to Random Forest.

VI. FUTURE SCOPE:

The UPI fraud detection system is effective, but there is still scope for improvement and addressing more challenges. Below are future enhancements:

Identified Problems:

Advanced Fraud Techniques: Fraudsters continuously evolve their strategies, making it challenging to detect new patterns.

Imbalanced Datasets:

Fraudulent transactions are rare compared to genuine ones, leading to skewed datasets and biased models.

Real-Time Processing:

The need for ultra-fast detection in real-time systems is critical.

Feature Enrichment: Current features may not capture all transaction behaviors effectively.

VII. CONCLUSION:

The UPI fraud detection project proposes a comprehensive approach to identifying and preventing fraudulent activities in real time, thereby enhancing the security of digital financial transactions. By leveraging a meticulously curated dataset, the project incorporates advanced analytical techniques to detect anomalies in transaction patterns with precision. It emphasizes the importance of analyzing critical transactional features, such as geolocation, transaction amount, and frequency, which play a pivotal role in identifying suspicious activities.

The methodology involves rigorous data preprocessing to ensure the quality and reliability of input data, followed by the development of a robust detection model capable of maintaining a delicate balance between minimizing false positives and maximizing detection accuracy. This enables the system to effectively differentiate between legitimate and fraudulent transactions.

The resulting solution is both scalable and efficient, capable of adapting to the growing volume and complexity of UPI transactions. By safeguarding users against financial fraud, the project not only protects individual users but also strengthens confidence in the digital payment ecosystem. Furthermore, it contributes significantly to reducing financial risks, supporting the sustainable growth of secure and trustworthy digital financial services.

XI. REFERENCES:

[1] J. Kavitha, G. Indira, Anil kumar, Shrinita, Bappan, ”FRAUD DETECTION IN UPI TRANSACTIONS USING ML”, IJRD, Article DOI: https://doi.org/10.36713/epra16459 DOI No: 10.36713/epra16459

[2] Yash Patil, Amar Shinde, Yash Parthe, Sameer Sayyad, “UPI FRAUD DETECTION USING MACHINE LEARNING”, irjmets, Volume:06/Issue:09/September-2024 Impact Factor- 8.187.

[3] Mainak Saha, Shaiban Mulla, Sumedh Gamre, “Fraud Detection In UPI Transaction Using AI”, IJCRT, © 2024 IJCRT | Volume 12, Issue 4 April 2024 | ISSN: 2320-2882,

[4] Miss. Sayalee S. Bodade, Prof. P.P. Pawade, “Implementation Paper on UPI Fraud Detection using Machine Learning”, JETIR, 2024 JETIR April 2024, Volume 11, Issue 4.

[5] Shabreshwari R M, Shafiya Mehrooz, Sidra Fatima, Tanmai R B, Prof. Ganesh Manasali ”UPI Fraud Detection Using Machine Learning”, IJEM, Volume 6, Issue 06 June 2024, pp: 98-100 www.ijaem.net ISSN: 2395-5252

[6] Sameer Kolekar, , Sourabh Panhale ,Dnyanendra Rengade ,Dipak Pawar, Prof. P.V,Kothawale “UPI fraud Detection Using Machine Learning”, IJSREM, Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

[7] Selvi P, Suryadharshan S “UPI Fraud Detection Using Machine Learning”, IJIRCCE, Volume 12, Issue 6, June 2024 , DOI: 10.15680/IJIRCCE.2023.1206055