UNVEILING HIDDEN PATTERNS: A DATA-DRIVEN APPROACH FOR DETECTING FRAUD IN ONLINE PAYMENTS

University at Buffalo, New York

Aditya Thakare*, Onkar Ramade+, Sourabh Kodag±
*athakare@buffalo.edu, +onkarraj@buffalo.edu, ±skodag@buffalo.edu

I. Abstract

Online payment fraud is growing in recent times, affecting businesses and consumers alike, as the increasing volume of these transactions opens newer avenues for malicious activities. Detecting fraud transactions in online payments is a complex challenge due to large volume, variety and velocity of the transaction data. This project displays a data-driven approach for detecting fraudulent transactions in online payment systems using machine learning and big data technologies. We propose a framework that leverages Hadoop for distributed storage and processing, HBase for data management, and machine learning models for fraud detection. Our approach aims to improve detection accuracy by identifying subtle patterns within transaction data that could indicate fraudulent behaviour. The proposed system is tested using a real-life dataset, and results show it can efficiently flag suspicious transactions, which in turn would offer the necessary toolkit to businesses for financial losses reduction and enhance transaction security.

II. Introduction

With the widespread adoption of online payment systems, the volume of digital transactions has skyrocketed, necessitating the implementation of powerful fraud detection and prevention systems. Online payment fraud is when individuals or organizations use payment systems for illicit objectives, resulting in financial losses, reputational damage, and legal consequences.

Machine learning techniques have gained significant attention in recent years as a potentially viable answer to this problem. Unlike traditional systems that rely on pre-defined rules, machine learning algorithms learn from previous transaction data, detecting complex patterns and anomalies that may signal fraudulent behaviour. Machine learning models can detect previously unknown types of fraud by using information such as transaction frequency, quantity, location, and user behaviour. However, the huge volume and variety of online payment data pose a challenge to these models, necessitating scalable and efficient infrastructure for data storage, processing, and analysis. This is where big data technologies like Hadoop and HBase come into play, providing a solution for efficiently managing and processing large amounts of data while maintaining detection capabilities.

In this project, we explore the use of a data-driven approach to online payment fraud detection, integrating machine learning with big data frameworks. We utilize Hadoop's distributed storage and processing capabilities to handle massive datasets, and HBase's real-time data management features to store transactional data. The goal is to create a scalable, efficient, and accurate fraud detection system capable of processing large amounts of transaction data in real time, allowing organizations to detect fraud as it occurs and take preventive action swiftly.

III. Dataset Description and Cleaning

In this project, the datasets used consists of approximately 1.5 million digital transactions that includes both legitimate and fraudulent transactions. The dataset captures various aspects of transaction behaviour, providing key features such as transaction amount, user information, product details, and the time and location of the transaction. These features are critical for detecting patterns indicative of fraudulent activity, such as unusual transaction amounts, high-frequency transactions, and abnormal user behaviours.

The dataset consists of several attributes, including but not limited to:

- Transaction Amount
- User Information
- Transaction Type
- Product Information
- Time and Location

This dataset allows for a thorough analysis of various transaction characteristics, providing the necessary foundation to develop and evaluate fraud detection models. The wide variety of features ensures that the models is trained to identify both overt and covert fraud patterns, enhancing the system's capacity to accurately identify possibly fraudulent transactions.

The dataset was properly cleaned to verify its integrity and suitability for analysis. Missing values were addressed by replacing numerical columns with the mean or eliminating rows with significant missing data. Duplicate entries were found and eliminated to ensure that each transaction was unique. Outliers, notably in transaction quantities, were identified and addressed to avoid distortion in the analysis. These pre-treatment methods guaranteed that the data was consistent, dependable, and suitable for model building.

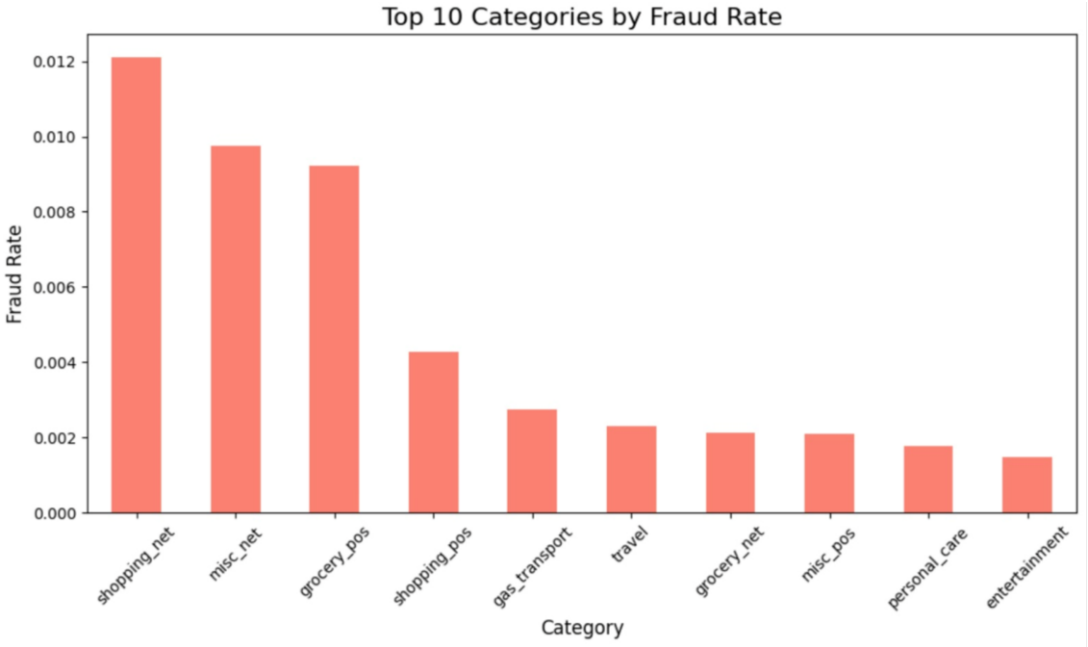

Fig 1: Distribution of Fraudulent transaction across product categories

IV. Machine Learning Models and Accuracies

This research sought to address six major questions, each tackled with an appropriate machine learning model to detect and analyse fraudulent transactions in the dataset. The methodology was based on a systematic approach where each question was approached with specific feature engineering, data pre-processing, and model selection strategies. The first step was data preparation, which included cleaning the dataset for missing values, removing duplicates, and applying techniques such as normalization and encoding to prepare the data for machine learning. For each of the six research questions, different models were implemented in order to explore different aspects of fraud detection.

➢ Decision Trees for classifying Fraud Transactions using Account Age

We analysed the relationship between customer account age and the likelihood of fraudulent transactions by investigating how account age, in conjunction with transaction amount, contributes to fraud detection using a Decision Tree classifier. The dataset was first balanced through oversampling of the minority class, which consisted of fraudulent transactions, to address class imbalance. Subsequently, the Decision Tree model was trained and evaluated with an accuracy of 80%, having a precision of 0.79 and recall of 0.81. These results highlight the significant role of account age in distinguishing fraudulent transactions from legitimate ones. The model extracted patterns related to account age and transaction behaviour and demonstrated that account age could serve as an important basis for fraud detection, informing future fraud prevention systems. This finding suggests that businesses could benefit from considering account age as a key feature when developing models to identify potentially fraudulent activities.

➢ Analysing Fraud Detection Patterns using k-Nearest Neighbours and Customer Age

We also tested another hypothesis in this regard: there is an increased possibility of fraudulent transactions coming from younger customers' age segments, aged 25-45 years. The k-NN showed better performance compared with Gradient Boosting, recording an overall accuracy of 86%, and an AUC as high as 0.92 while classifying fraudulent and valid transactions in the test datasets. These results indeed suggest that k-NN captures age-related patterns in fraud detection. The model was trained with neighbours set to 5, striking a balance between model complexity and stability. We applied SMOTE to handle class imbalance, which helped to improve the fraudulent transaction classification capability of our model. The recall of the k-NN model for fraudulent transactions is 0.91, which implies that it is very efficient at catching fraud with minimal false negatives. In contrast, Gradient Boosting performed worse with an accuracy of 75% and AUC of 0.80. These findings support our hypothesis that younger customers may exhibit patterns that are more likely to be associated with fraudulent activities. The high recall and AUC reinforce k-NN's capability to detect fraud and validate the hypothesis about age-related behaviour in fraud detection.

➢ Exploring Time-Based Patterns in Fraudulent Transactions Using CatBoost

The analysis aimed to explore whether fraudulent transactions differ by hour, assuming that the time variable could be crucial for classifying fraud. In this way, it suggests that there are specific moments in a day that can potentially provide more fraud and, hence, understanding time-of-the-day patterns of fraud might help organizations in detecting and deterring fraud more effectively. This model was a result of CatBoost; its precision is 0.84, while recall is 0.79 for fraudulent transactions.

Thus, having an overall accuracy of 82%, this model surely catches time-related patterns from the dataset. In the report classification, its weighted F1-score shows up to 0.82, which presents its capability of identifying fraudulent activities at certain hours. These results validate that time-based patterns could serve as an important axis in fraud detection, helping organizations to better target high-risk time periods in mitigating fraud.

➢ Evaluating Geographic Factors in Fraud Prediction Using Gradient Boosting

This analysis aimed to explore whether geographic distance, cardholder demographics, and transaction details can predict fraudulent transactions. Fraudulent activities often occur from locations that are far away from the cardholder's usual area or in densely populated regions where fraud is more prevalent. Demographic factors such as gender, age, and job title may provide insight into targeted groups, while transaction-specific details like amount and merchant type can also help identify suspicious activities. The geographic distance between the cardholder and merchant was computed using the Haversine formula. By combining these geographic and demographic features with transaction data, a Gradient Boosting model was used to achieve an accuracy of 99.78%. The model also returned a very high ROC-AUC score of 0.947, indicating that it was effective in detecting fraudulent transactions. Though class imbalance resulted in a recall rate of fraudulent transactions that is quite low, the overall model performance indicates how important this incorporation of factors like location, demographics, and attribute-level transaction information could be towards enhanced fraud detection.

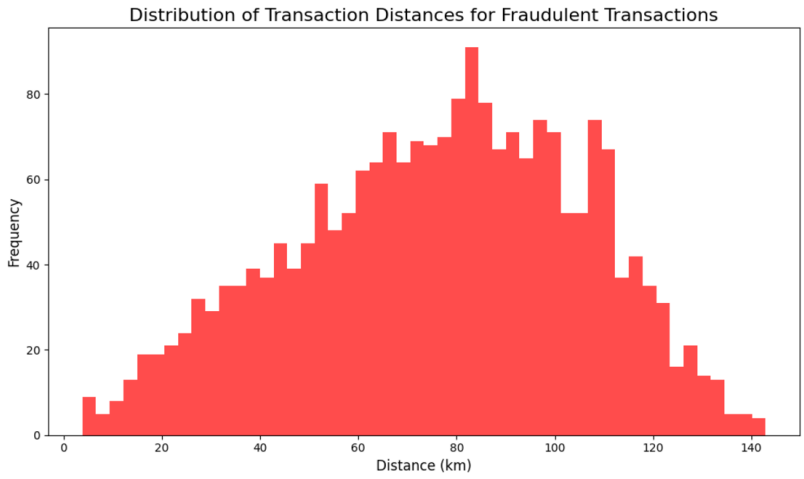

Fig 2: Distribution of Distances Between Fraudulent Transaction Locations and Customer Locations

➢ Identifying Fraudulent Transactions Based on Product Categories Using Random Forest

This hypothesis suggests that certain product categories are more likely to experience fraudulent transactions than others. Fraudsters generally target high-value or high-demand products because these offer higher margins and are also easier to resell. Products like electronics or luxury goods are particularly vulnerable because of their desirability and high cost. Transaction data, when analysed regarding product categories, can develop patterns that will identify fraud risks higher for certain sectors. This concept, therefore, calls for the classification of transactions based on product type in order to understand their correlation with fraud, which would lead to more accurate fraud detection models. These models would provide businesses with the ability to focus on high-risk categories, thus improving their capabilities in fraud detection and thereby reducing associated losses. In this analysis, the Random Forest model was used, which had a very good

performance with an accuracy of 1.00 and very good precision and recall, reflecting its efficiency in identifying fraud in product categories.

- Fraud Detection Based on Geographic Location and Product Categories Using XGBoost

In this analysis, we explore the hypothesis on fraudulent transactions that might set up specific patterns based on geographic locations and proximity to metro cities, besides the susceptibility of fraudulent activities within some product categories. Fraudsters may target areas with dense populations or near major urban centres, where anonymity and the volume of transactions make fraud detection more challenging. Likewise, high-demand or high-value product categories, like electronics or luxury goods, also form the target of many fraudsters because of the associated high margin of profit and easy resale value. The study will look for correlations of transactions with both geographic location and product categories for fraud occurrence. XGBoost yielded an excellent result on the model, with an accuracy of 0.9991 and a ROC AUC score of 0.9981, showing that the model would correctly identify fraudulent transactions. The classification report indicated high

precision and recall, especially for fraudulent transaction classes, which proved that geographical and product-related factors do play important roles in fraud detection.

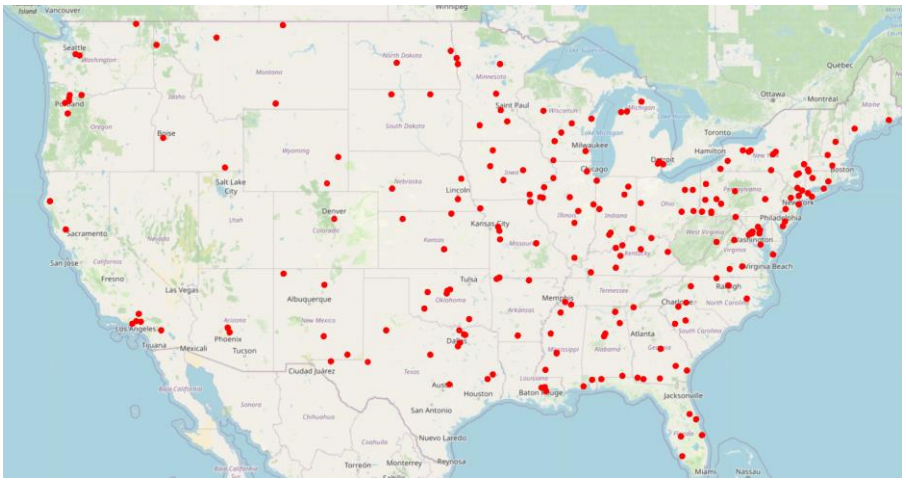

Fig 3: Map of the USA Highlighting the Region with the Highest Fraudulent Activity

V. Results

The factors that affect the probability of fraud turn out to be significant drivers and involve customer demographics, transaction details, and geographic features. All the models, especially XGBoost and k-NN, performed well and showed accuracy rates of over 99% in some cases. Interestingly, customer age, time of day, distance from the usual location in a geographic sense, and product categories were among the influential factors in the accuracy of fraud detection. On further analysis, it was evident that customers who were younger and closer to a metro city tended to commit fraud. Another feature was high-value product categories that generally had a higher rate of fraud.

Other methodologies, like SMOTE, helped increase the minority class and consequently better trained the models to detect actual fraudulent cases without getting over-optimistic toward the larger fraction of classes. These studies underpin the fact that fraud detection should be oriented in many dimensions: first through a transactional perspective but, preferably added by demographic

and location-based angles. These models will assist in honing fraud prevention strategies and enhancing fraud detecting accuracy.

VI. Future Works

While the current model for fraud detection and its web application have shown promising results, there is a lot of scope for its future improvement and expansion. One of the key directions in which future work could continue is in the exploration of more advanced machine learning models. They could be fine-tuned to better capture the complex, nonlinear relationships that define fraudulent activities. Besides, hyperparameter optimization may further improve the performance of these models and, by extension, fraud detection.

In terms of the web application, the current system was successfully hosted locally, displaying results and allowing users to interact with the dataset. However, to make the application more scalable and accessible, future work could focus on deploying the application to cloud platforms like AWS or Google Cloud, making it globally accessible and with better resource management. Other improvements would include real-time fraud detection integrated into the web application; it would permit the system to show results in a live manner, updating on every transaction. This may also greatly enhance user experience and make the application more useful to businesses who seek to apply fraud detection to their systems for detecting fraud in payment systems.

The interpretability of decisions by the model could be enhanced in further work. Understanding the rationale of a fraud prediction is important, especially in a real-world scenario. Techniques such as SHAP or LIME might be used to explain the model's predictions, building trust and understanding of the decision-making process among stakeholders. Finally, increasing the dataset through the inclusion of more diverse data sources, such as user demographics or external datasets on payment patterns, would enhance the robustness and accuracy of the model.

VII. Conclusion

Finally, the studies and models established for fraud detection have provided useful insights into the fundamental aspects that contribute to fraudulent transactions. The study demonstrated the efficiency of merging transactional data, consumer demographics, and geographic information for fraud detection using a range of machine learning techniques such as k-NN, XGBoost, and Random Forest. The findings emphasize the necessity of evaluating both global and micro-level characteristics, such as consumer age, product category, and geography, when detecting possible fraud. Furthermore, approaches such as SMOTE helped balance the dataset, resulting in robust model performance. This study lays the framework for more precise and comprehensive fraud detection systems that can adapt to changing fraudulent practices and deliver actionable information to businesses and financial institutions.

VIII. References

[1] https://www.databricks.com/blog/2021/04/13/identifying-financial-fraud-with-geospatial-clustering.html

[2] https://trustfull.com/articles/how-email-age-unmasks-new-account-fraudsters