Mining Transactional Data to Combat Fraudulent Activities and Promote Digital Literacy

Akanksha Tyagi
Data Analytics
SJSU
San Jose, USA
akanksha.tyagi@sjsu.edu

Anjali Himanshu Ojha
Data Analytics
SJSU
San Jose, USA
anjalihimanshu.ojha@sjsu.edu

Sakshi Manish Mukkirwar
Data Analytics
SJSU
San Jose, USA
sakshimanish.mukkirwar@sjsu.edu

Srushti Lalit Doshi
Data Analytics
SJSU
San Jose, USA
srushtilalit.doshi@sjsu.edu

***Abstract*—In the modern digital era, online transactions constitute the majority of businesses' revenue. As such, online fraud poses a huge risk often resulting in financial losses and operational setbacks. Through the application of advanced data mining techniques, this project aims to identify the patterns, trends, and anomalies that are indicative of fraudulent activities. The project analyses patterns in the IEEE-CIS Fraud Detection dataset concerning features like transactional amounts, time of transactions, distance of transactions, and email addresses. The results demonstrate how businesses can utilize online transactional data to proactively safeguard their operations and foster a more secure financial environment. In addition, these insights can also inculcate digital literacy among consumers.**

***Index Terms*—data mining, fraud detection, and prevention, machine learning**

I. Introduction

Fraudulent transactions remain a critical challenge for small e-commerce platforms, often leading to revenue loss, charge-backs, and damaged reputation. Early detection of the symptoms of fraudulent activities enables small businesses to safeguard their operations. This project aims to highlight the indicators of fraudulent activities during online transactions using advanced data mining techniques. The project uses a large-scale IEEE-CIS Fraud Detection dataset [1] which consists of a diverse set of features. The data also makes sure that the privacy of the transactions isn't compromised by providing anonymized features.

A. Motivation or Justification

E-commerce has indeed reformed the retail landscape and opened a wide scope for business opportunities; on the other hand, it brings new risks, which come in forms such as fraudulent transactions. Online fraud has escalated significantly in recent studies and disproportionately affects small to medium-sized enterprises (SMEs) due to their limited resources to invest in sophisticated fraud detection systems. Thus, this work was motivated by an urgent need for publicly available and applicable fraud-detecting capabilities that can be enacted by smaller e-commerce systems to defend both themselves and their clientele. The National Cyber Security Alliance (NCSA, 2019) Small Business Cybersecurity Report [2] mentions that many small businesses fold within months of falling victim to a major fraud. Better fraud detection can prevent those losses, allowing companies to work in a stable environment. Accordingly, the project has been designed to develop a model using the IEEE-CIS Fraud Detection dataset [1], to identify fraud with efficacy through the use of advanced data mining techniques, thereby reducing risks and making the e-commerce world even safer.

B. Importance and Contribution to Community

This project is crucial because it has a critical role in enhancing the resiliency of small and medium-sized businesses and e-commerce platforms from the increasing threat of online fraud. This project saves these businesses from substantial financial losses with sophisticated data mining to uncover subtle aberrations and suspicious patterns in transaction data. It also protects personal and financial information, which further ensures customer confidence. It democratizes access to fraud detection technologies, which were hitherto afforded only by larger corporations that have bigger resources at their command, therefore evening out the playing field for smaller players.

1) Key Contributions of the Project: The project contributes in the following ways.

a) Advanced Detection Techniques: Provides state-of-the-art data mining techniques to detect and notify a business in advance of fraudulent activities before they result in bottom-line losses.

b) Scalable and Adaptable Solutions: The project contributes towards providing scalable solutions that can be adapted to various business sizes and types, which will help in broad applicability and ease of integration.

c) Economic Impact: The reduction in financial losses due to fraud will lead to sustainable operations and growth for the business.

d) Accessibility of Technology: It offers small businesses easy-to-implement and affordable fraud detection tools which will create a safer and more seamless transaction environment for all users.

II. Related Works and Research

The detection of fraudulent activities is becoming increasingly harder with advancements in technology. Thus, rule-based systems are no longer sufficient leading to the demand

for advanced ML-based data mining techniques [3], [4]. Modern data mining techniques involving supervised and unsupervised algorithms provide scalable and robust approaches to analyzing large datasets for anomalies [5]. Machine learning models like decision trees, random forests, and Support Vector Machines have demonstrated great success in the detection of fraudulent activities in the financial sector [6]. Notably, a hybrid approach that combines various machine learning techniques such as Random Forests, Gradient Boosting, and Logistic Regression has proven effective. These methods excel in detecting fraud by analyzing inconsistencies in transaction data and learning from both fraudulent and legitimate transaction patterns [7].

In addition to machine learning models, digital literacy plays a pivotal role in mitigating fraud by enabling consumers to recognize fraudulent patterns. Such patterns can be gleaned from the datasets of transactions. Existing studies have shown that vigilant consumers can help in combating online fraud to a great extent [8]. Combining advanced data mining techniques with digital literacy initiatives creates a robust framework for tackling fraudulent activities. Using transactional data for real-time fraud detection, alongside efforts to enhance digital awareness, fosters a more secure digital landscape.

III. Data Preprocessing

The initial phase of our project involved data preprocessing steps to ensure the quality and usability of the dataset for effective fraud detection analysis. The dataset is provided in two separate files *transaction* and *identity*. The process began with the integration of these distinct datasets, where we merged them based on their transaction IDs. We did this merge to combine relevant information scattered across both datasets. The dataset contains 434 features. Some of these features are anonymized to preserve privacy. These features show up *V*, *M*, and *C* features with an index suffix. For example, *Vxxx* are features like ranking, counting, and other entity relations.

1) Data Integration: We formed a unified dataset that includes all the necessary attributes for each transaction by joining the two datasets on the transaction ID. This step was important in providing an analytic platform where each transaction can be analyzed in detail.

2) Handling Missing Values: The data have 434 different features, first, we drop the columns with more than 90% of missing values. For other columns with missing values, we used the Mean Value or Median value for continuous features and the most frequent value for categorical features.

3) Data Transformation: The date column is given as the relative integer values in seconds, we convert those to days and create a relative ordering for time-series. There are several columns with string type values for *T* and *F* values, we converted them to 1, 0 values.

4) Outlier Analysis: During our investigation of TransactionAmt, we discovered some extreme outliers using the Interquartile Range (IQR) method. These were transactions that stood out because they were much higher than the quartile range. We removed these outliers to prevent them from distorting our analysis. This step helped make sure that our fraud detection efforts are based on genuine patterns, making our findings more reliable and our models more effective.

5) Derived Features: We also derived multiple new features like P_emaildomain and R_emaildomain generalization by assigning high-level domains like 'yahoo.com', 'yahoo.com.mx', and 'yahoo.co.uk' resolved to yahoo. We use a similar technique for the browser version categorization. We also generate features like time of TransactionHour, TransactionDayOfWeek, etc.

6) Feature Encoding: In the data preprocessing phase of the project, categorical variables such as ProductCD, card4, DeviceType, Purchaser email domain, and Recipient email domain were transformed into numerical formats suitable for machine learning models. One-hot encoding was applied to variables with no ordinal relationship, creating binary columns for each category. Label encoding was utilized for the Purchaser email domain, converting each domain into a unique integer. Additionally, frequency encoding was used for the recipient email domain to encode categories based on their occurrence frequency, highlighting the prevalence of each domain within the data.

IV. Data Analysis

This section provides a detailed analysis of the patterns found in the IEEE-CIS Fraud Detection dataset [1]. Expected trends and patterns for fraudulent activities are analyzed in section IV-A while hidden patterns, anomalies, and strange structures are analyzed in section IV-B.

A. Expected Trends and Patterns

1) Class Distribution: Due to the nature of the fraudulent activities, it is expected that the fraudulent cases are less frequent than non-fraudulent cases. Thus, the dataset is expected to have a high-class imbalance as shown in figure 1.

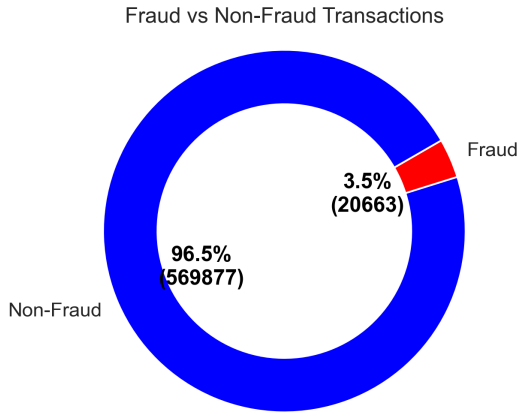

Fig. 1. Distribution of classes in the dataset showing high-class imbalance

Transaction data is time series data, where different cardholders make different purchases over time. As we can see in the figure 2 there is no overlap between train and test

data. There is a gap between train and test data to ensure that the model is evaluated on completely unseen future data, preventing ”data leakage” where the model might learn patterns from the test set.

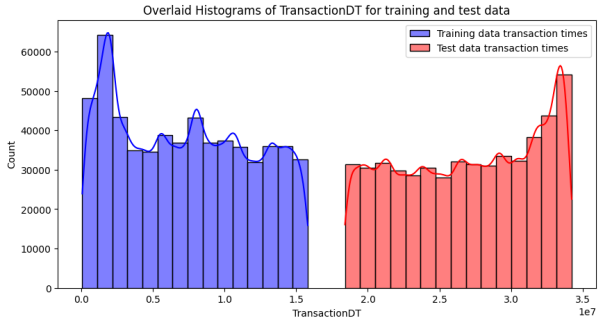

Fig. 2. Train and Test data over time.

2) Distribution of Transaction amounts: The fraudulent transactions are expected to have high mean and variance because fraudsters want to maximize their gains and avoid detection respectively. Figure 3 shows this pattern via a box plot. We also notice outliers in the transaction amounts. Note that the y-axis is in log scale to accommodate all the transaction amounts. Thus, the outliers are quite high and large in value.

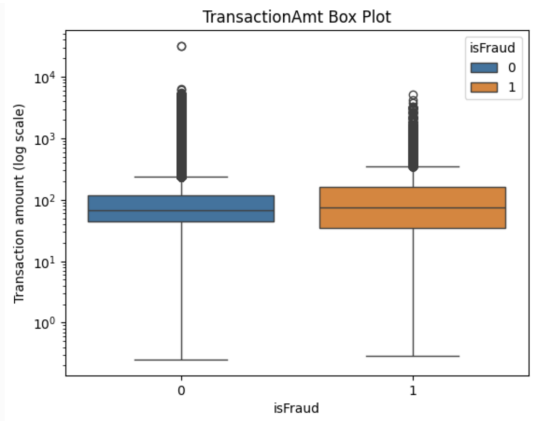

Fig. 3. Box plot of transaction amounts showing that fraudulent transactions have a higher mean and variance.

3) Distance Analysis for Transactions from billing address: Figure 4, shows the scatter plot illustrating the relationship between dist1 (distance from billing address) and TransactionAmt (transaction amount), with non-fraud transactions in blue and fraud transactions in red. The clustering of non-fraud transactions at short distances and lower amounts aligns with typical consumer behavior: people generally make higher-value purchases closer to familiar locations, like their homes or workplaces, where they feel secure. Some fraudulent transactions are appearing with higher transaction amounts at higher distances. This pattern indicates that when high-value transactions occur far from the billing address, they deviate from usual consumer behavior and could signal fraud, as legitimate high-value transactions are more likely to be local.

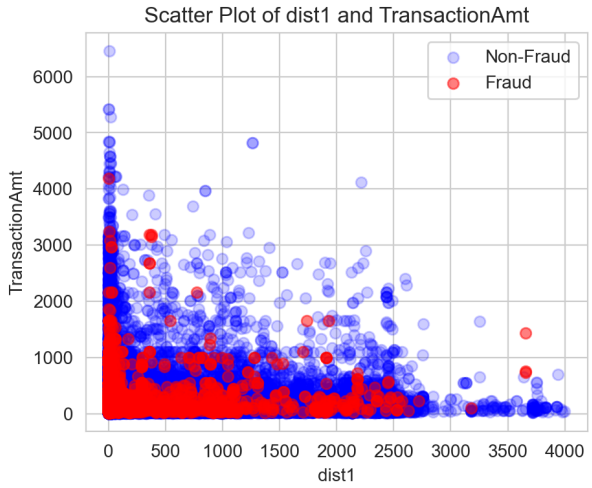

Fig. 4. This plot shows the data distribution for the Transaction Amount and distance from the billing address. It is seen that most of the higher amount transactions happen nearby, but as the distance increases there is an increase in the average transaction amount for the fraudulent transactions.

B. Hidden Patterns, Anomalies, and Strange Underlying Structures

1) Hourly Variations in Fraudulent Activities: The dataset provides a relative timestamp for each transaction. The timestamp is converted into the relative hour of the day. The number of transactions and fraud proportion can now be computed for each hour of the day over the entire time duration of the dataset. Figure 5 shows that the number of transactions is high on certain hours and low on other hours. It is noticeable that the high proportion of frauds coincides with a low number of transactions showing that fraudulent activities are much more prevalent during late hours when the number of transactions is also fewer. It contradicts the usual understanding that frauds happen more among high transactions so they go unnoticed by businesses.

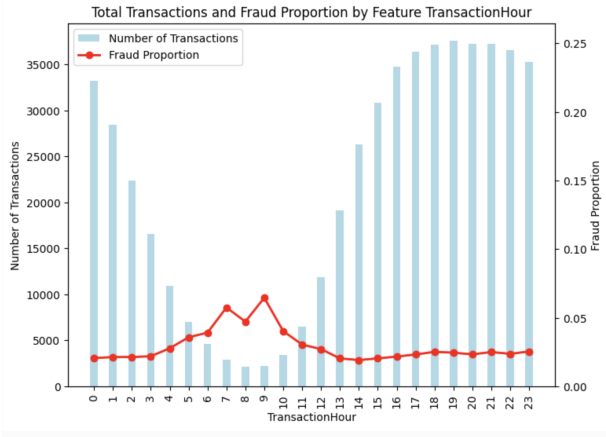

Fig. 5. The proportion of frauds is higher on certain hours which coincides with a lower number of transactions.

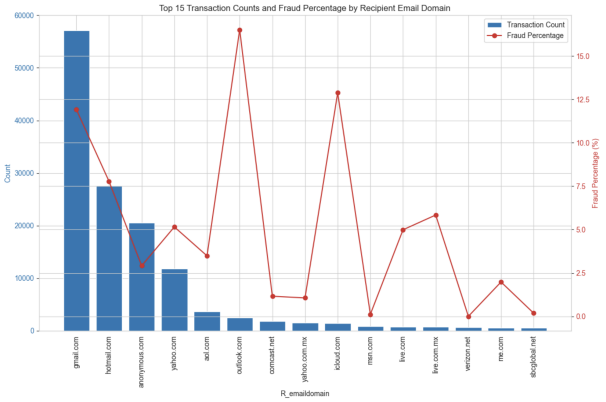

Fig. 7. Top 15 transaction counts and fraud percentage by recipient email domain showing high fraud rates for icloud.com, outlook.com, and gmail.com.

2) Analysis of Transaction Volume and Fraud Incidence Across Email Domains: Figure 6 and 7 illustrate transaction counts and fraud percentages for the top 15 email domains, showcasing notable variations in fraud risk. For recipient domains, peaks in fraud percentages at *hotmail.com* and *anonymous.com* suggest vulnerabilities or targeted attacks. Similarly, the purchaser email domain data indicates a high volume of transactions at *gmail.com*, but elevated fraud percentages at *hotmail.com* highlighting the need for strict monitoring. These insights are critical for tailoring fraud prevention strategies to specific email domains, enhancing overall transaction security. Contradictory to common knowledge, we notice high fraud percentage for domains like *outlook.com*, *icloud.com*, and *gmail.com* for receiver. These domains are owned by technology giants like Microsoft, Apple, and Google respectively. One possible explanation is that fraudsters use these domains so the consumers are easily duped.

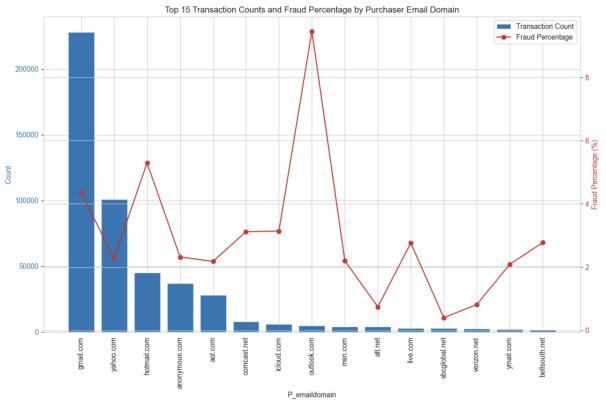

Fig. 6. Top 15 transaction counts and fraud percentage by purchaser email domain showing high fraud rate for outlook.com, and hotmail.com.

3) Transaction Counts and Fraud Percentage by Weekdays: Analyzing transaction patterns by weekdays is crucial for identifying trends that can help enhance fraud detection strategies. 8 illustrates the distribution of transaction counts and fraud percentages by weekdays, revealing a noticeable increase in

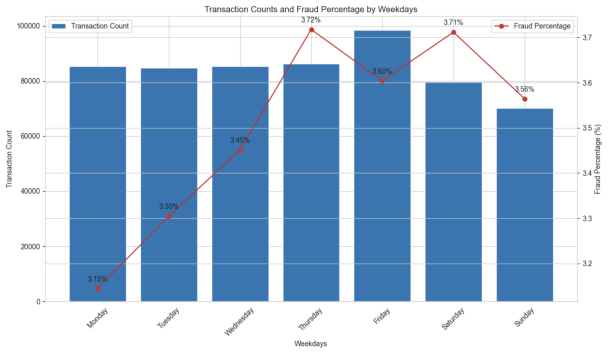

Fig. 8. Transaction counts and fraud percentage by weekdays showing high fraud rates around weekends.

both metrics towards the weekend. Monday has the lowest fraud percentage (3.15%) and moderate transaction count. The chart indicates that fraud rates tend to increase towards the end of the workweek, particularly on Thursdays and Fridays, despite relatively steady transaction volumes.

4) Transaction Counts and Fraud Percentage for different Processors and Card Type: Figure 9 presents two heatmaps comparing total transaction counts and fraud percentages by processor and card_type (card6) combinations. The left heatmap shows that *Visa* and *Mastercard* debit cards have the highest transaction counts. Notably with *Visa* debit transactions reach 301,023. It can be noticed that *Discover* credit cards have the highest percentage of fraud.

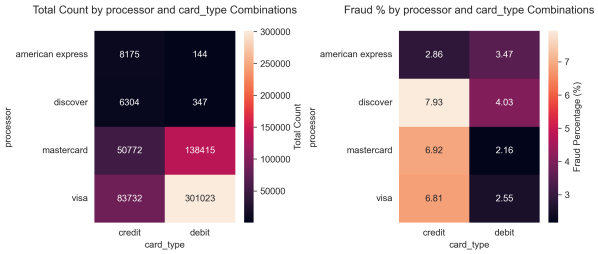

Fig. 9. Number of transactions and fraud percentage for different card companies and card types showing highest fraud percentage for Discover credit cards, but Credit cards have more number of frauds compared to the Debit cards. Debit cards have an added layer of security like 4-digit PIN, which can be helpful to mitigate the transactions from stolen cards.

In the right heatmap, fraud percentages are highest for *Discover* credit transactions (7.93%), with notable fraud rates also for *Mastercard* credit (6.92%) and *Visa* credit (6.81%). Debit transactions across all processors generally exhibit lower fraud percentages, suggesting that fraud is more prevalent in credit transactions, particularly for *Discover* cards. This highlights a pattern where credit transactions, especially with Discover are more susceptible to fraud than debit transactions. An unexpected observation is that American Express's debit card has a higher fraud rate than American Express's credit card. It contradicts the popular opinion that credit cards are more prone to fraud.

5) Fraud Transaction distribution with respect to distance from Billing addresses Comparisons: In Figure 10, the butterfly plot compares the distribution of fraud and non-fraud transactions based on the dist1 variable, which likely represents transaction distance from the billing address.

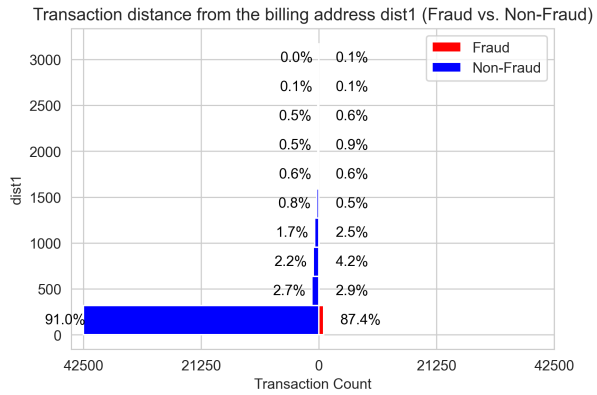

Fig. 10. Different transaction distances from the billing address are observed. Most transactions occur in nearby locations, resulting in a skewed data distribution. However, as the transaction distance increases, the likelihood of the transaction being fraudulent also increases.

Non-fraud transactions (in blue) are highly concentrated near zero distance, with **91%** of them occurring close to the billing address, indicating that legitimate purchases are typically local. In contrast, fraud transactions (in red) are a bit more spread across larger distances, but still have the highest transactions occurring close to the billing address. This pattern suggests that larger transaction distances could be an indicator of potential fraud, as fraudulent transactions tend to deviate from the typical proximity behavior observed in legitimate transactions, but it's not a very strong predictor.

6) Variations in Transactions and Frauds with respect to Device Type: Figure 11 shows that desktop devices have the highest number of transactions but the percentage of frauds is more on mobile devices. Thus, consumers and businesses need to be extra careful while conducting online transactions via mobile devices.

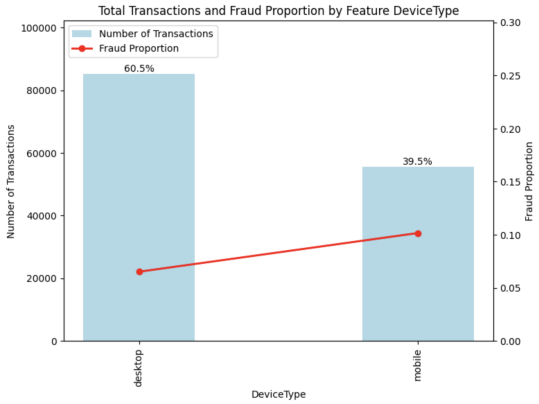

Fig. 11. Variation of transactions and frauds with respect to device type showing high fraud percentage on mobile devices.

7) Fraud Transactions with different browsers and further analysis for Google Chrome browser: We analyzed different browsers used for transactions (Figure 12) and found that mobile device browsers have higher fraud rates. Further investigation into Chrome, the most popular browser (Figure 13), reveals that older versions and mobile devices exhibit higher fraud rates.

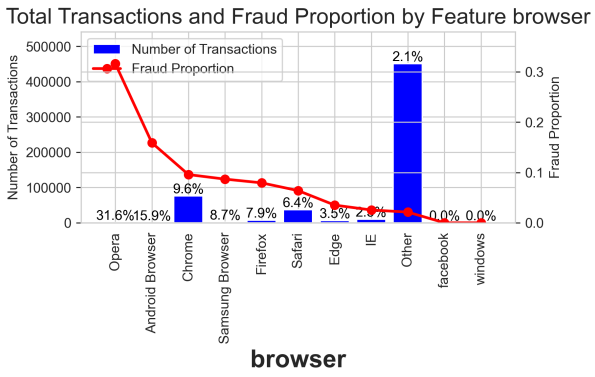

Fig. 12. Total transactions and Fraud transactions for different browsers in different devices. As we can see there are a lot transactions happening on other platforms, but Chrome has the highest fraud rate among the popular browsers. Figure 13 shows the further breakdown for different versions

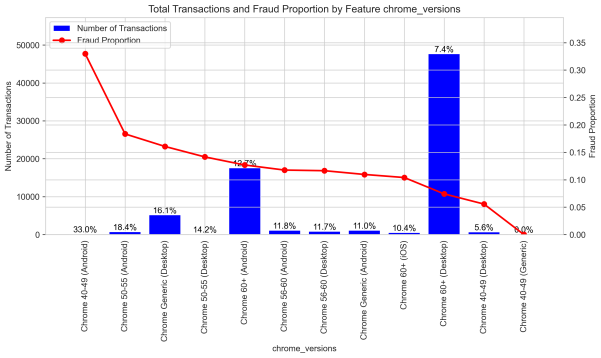

Fig. 13. Fraud transaction with a version of Chrome browser on a different platform. It shows that there is a higher number of frauds for mobile browsers and lower versions of desktop Chrome. And the newer version of Chrome browsers has better security features, so they have less fraud.

8) Feature Selection and Correlation Matrix: The dataset has a large number of features so training any Machine Learning model will be very challenging, so it's important we use a subset of features that captures the variance in data and we can use these to build a model. In this process we created a User and aggregated features like mean and deviations of amounts and number of days and time of transaction etc. We added 28 different features (TransactionAmt_card1_std, D9_card1_mean, etc). Figure 14 shows the top 25 important features. Figure 15 shows the correlation between top features, and it shows the similar types of features like Vesta features (V*) and counting related features (C1-C14) show the higher correlation between them.

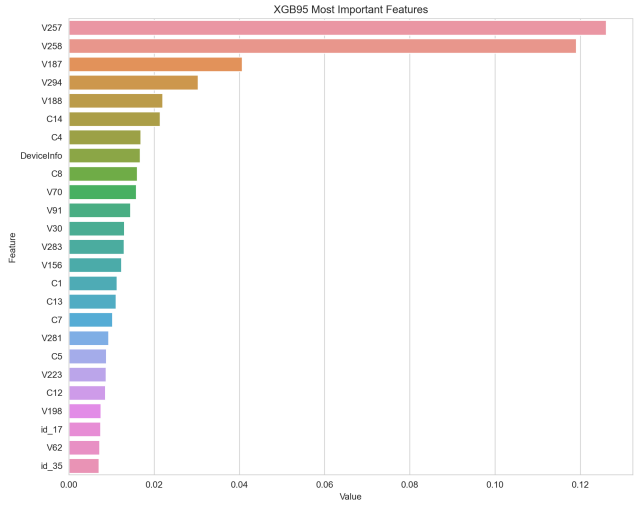

Fig. 14. Top 25 Important Features extracted using XG-Boost method [9].

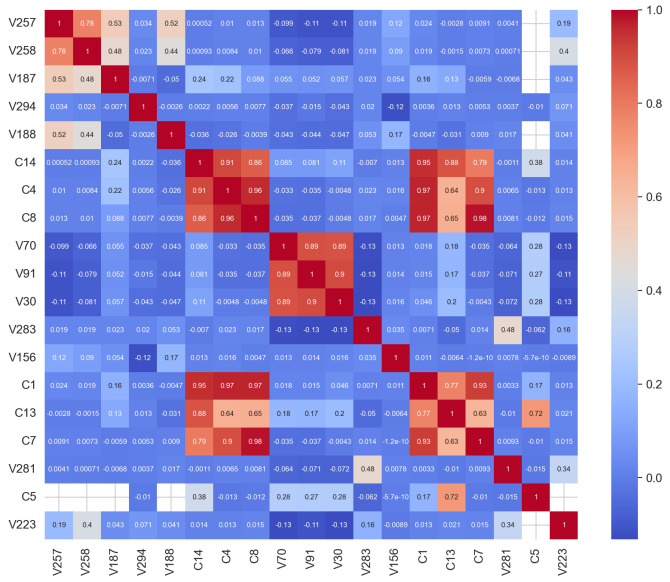

Fig. 15. Correlation Matrix for the Top 20 features excluding the DeviceInfo variable as it's a categorical variable with multiple values.

V. Community Contribution

This project aimed to provide safeguarding mechanisms to both consumers and small businesses against online transaction frauds through insights gained from data mining. Frauds lead to financial repercussions that affect everyone including business owners, employees, and consumers. Such setbacks have a ripple effect that can also lead to layoffs and business closures destabilizing the economy. Businesses that proactively take measures to avoid fraud are trusted more by the consumer encouraging their support. Therefore, by minimizing fraudulent activities, businesses can strengthen themselves financially, retain employees, maintain fair wages, and focus on sustainable growth.

Credit card fraud detection is essential not only for preventing financial losses but also for ensuring that genuine customers are not inconvenienced. Based on the analysis done in section IV, we provide guidelines to help reduce online fraud. As we see figure 5 and 8, there are certain times when fraudulent activities are more active. Thus, it is recommended to be extra careful during these times. Figure 11 shows mobile devices are more prone to fraud. So, consumers should be extra careful while conducting transactions especially paying extra attention to email addresses since certain email addresses are highly attached to frauds as shown in figure 6 and 7. Figure 12 shows that mobile devices have higher percentages of fraud, and 13 shows for well-known browsers like Chrome the older versions have relatively higher fraud rates compared to the updated versions. Thus, it is recommended to update the browsers frequently so that the security updates are installed.

VI. Conclusion

This project analyzed the trends in online transactional data in the IEEE-CIS Fraud Detection dataset. We identified patterns and indicators associated with online fraud. We offer practical insights into transactional behavior, device types, email domains, and specific time periods that show heightened risks of fraud. We found out that there's a high risk of fraud on mobile devices compared to desktop devices. In addition, there's high fraudulent activity during specific hours (late-night transactions) when overall transaction volume is lower. Thus businesses should prioritize enhanced monitoring of mobile transactions and consider stronger measures during these high-risk hours. These insights are pivotal in enabling businesses and consumers to design fraud prevention strategies that prevent financial losses and enforce trust with consumers in an increasingly competitive digital world.

References

[1] A. Howard, B. Bouchon-Meunier, I. CIS, inversion, J. Lei, Lynn@Vesta, Marcus2010, and P. H. Abbass, "Ieee-cis fraud detection," https://kaggle.com/competitions/ieee-fraud-detection, 2019, kaggle.

[2] National Cyber Security Alliance, "Small Business Cybersecurity Report," https://www.staysafeonline.org, 2019, [Online; accessed 30-October-2024].

[3] L. Cao, "Ai in finance: challenges, techniques, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–38, 2022.

[4] A. Ali, S. Abd Razak, S. H. Othman, T. A. E. Eisa, A. Al-Dhaqm, M. Nasser, T. Elhassan, H. Elshafie, and A. Saif, "Financial fraud detection based on machine learning: a systematic literature review," *Applied Sciences*, vol. 12, no. 19, p. 9637, 2022.

[5] X. Niu, L. Wang, and X. Yang, "A comparison study of credit card fraud detection: Supervised versus unsupervised," *arXiv preprint arXiv:1904.10604*, 2019.

[6] A. Jain and S. Shinde, "A comprehensive study of data mining-based financial fraud detection research," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*. IEEE, 2019, pp. 1–4.

[7] E. Malik, K. Khaw, B. Belaton, W. Wong, and X. Chew, "Credit card fraud detection using a new hybrid machine learning architecture. mathematics 10: 1480," 2022.

[8] P. Gomber, J.-A. Koch, and M. Siering, "Digital finance and fintech: current research and future research directions," *Journal of Business Economics*, vol. 87, pp. 537–580, 2017.

[9] C.-P. Hsieh, Y.-T. Chen, W.-K. Beh, and A.-Y. A. Wu, "Feature selection framework for xgboost based on electrodermal activity in stress detection," in *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*. IEEE, 2019, pp. 330–335.