SJSU SAN JOSÉ STATE UNIVERSITY

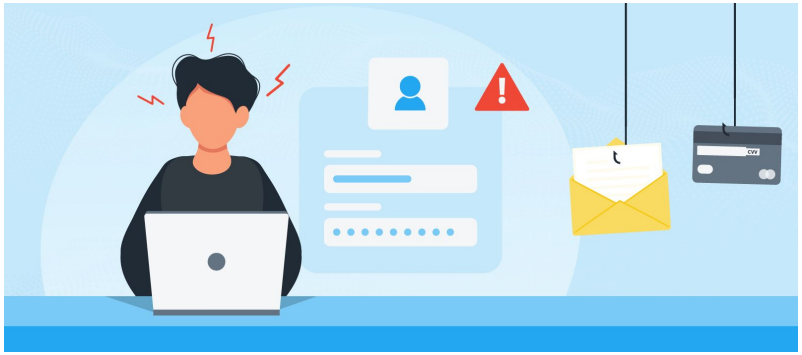
Mining Transactional Data to Combat Fraudulent Activities

Under the guidance of Dr. Taehee Jeong

Group 7

Anjali Ojha
Akanksha Tyagi
Srushti Doshi
Sakshi Mukkirwar

Agenda



1. Motivation
2. Dataset Overview
3. Data Mining and Analysis
4. Data Preprocessing
5. Feature Engineering
6. Modelling
7. Technical Novelty
8. Key Findings and Insights
9. Summary
10. References

Motivation

- **Increasing Global Fraud** : According to a recent report by the Federal Trade Commission, consumers lost over $10 billion to fraud in 2023, marking a significant 14% increase from the previous year.
- Escalating fraud techniques outpace traditional security measures, necessitating more sophisticated solutions.
- **Project Goal:** Understanding hidden patterns and fraud predictions for consumers and small businesses against online transaction fraud using data mining insights.

As Nationwide Fraud Losses Top $10 Billion in 2023, FTC Steps Up Efforts to Protect the Public

Investment scams lead in reported losses at more than $4.6 billion

February 9, 2024

Tags: Consumer Protection | Bureau of Consumer Protection | Consumer Sentinel Network

Newly released Federal Trade Commission data show that consumers reported losing more than $10 billion to fraud in 2023, marking the first time that fraud losses have reached that benchmark. This marks a 14% increase over reported losses in 2022.

Consumers reported losing more money to investment scams—more than $4.6 billion—than any other category in 2023. That amount represents a 21% increase over 2022. The second highest reported loss amount came from imposter scams, with losses of nearly $2.7 billion reported. In 2023, consumers reported losing more money to bank transfers and cryptocurrency than all other methods combined.

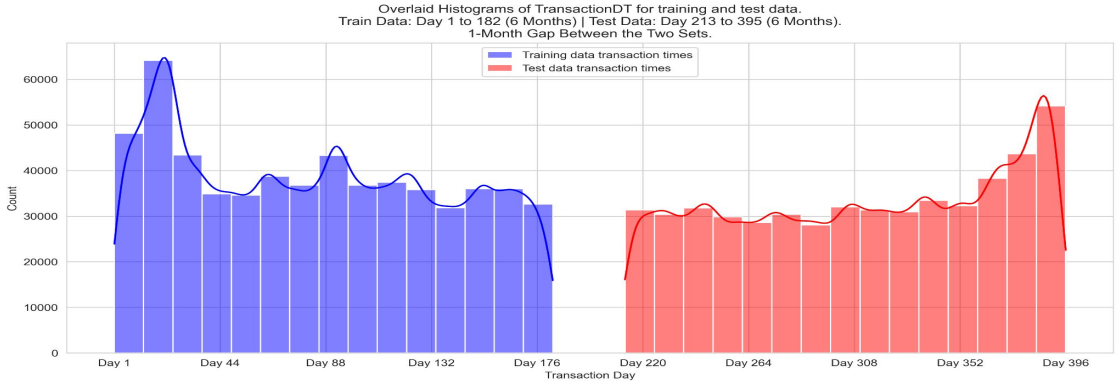
https://www.ftc.gov/news-events/news/press-releases/2024/02/nationwide-fraud-losses-top-10-billion-2023-ftc-steps-efforts-protect-public

Dataset Overview : IEEE-CIS Fraud Detection

SJSU SAN JOSÉ STATE UNIVERSITY

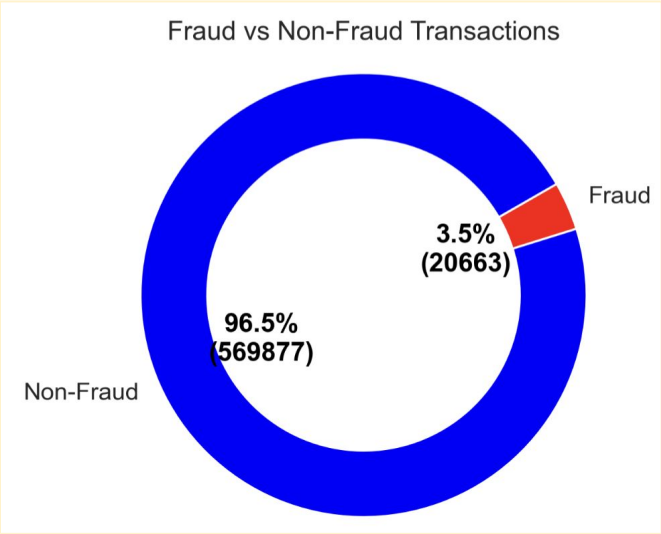
	Train	Test	Details
Transaction Features = 393	train_trainsaction.csv samples = 590,540	test_trainsaction.csv samples = 506,691	contains transaction-related features, including *transaction amounts, timestamps, product-code, card-type, email-domains, address, region, and fraud labels.*
Identity Features = 40+1	train_identity.csv samples = 144,233	test_identity.csv samples = 141,907	includes identity-related information, like *device type, browser information, operating-system, and anonymized personal details.*

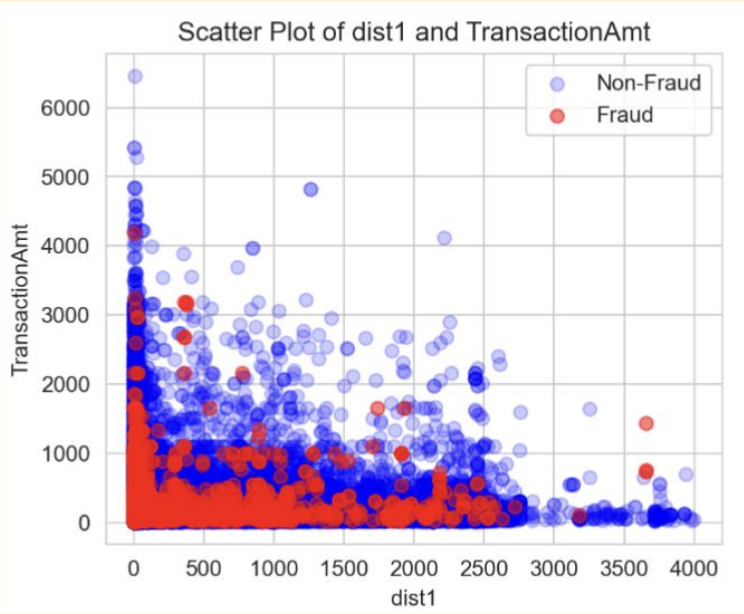
Note *- The training and test sets are split across transaction and identity files, with option to merge these on a* ***TransactionID*** *field.*
Our target variable is ***isFraud*** *column.*



Data Mining and Analysis

Expected Trends and Patterns [1]





> This chart highlights the class imbalance in the dataset, where fraudulent transactions constitute only **3.5% of the total**, underscoring the rarity and difficulty in detecting fraud.

> The scatter plot reveals most legitimate transactions occur close to billing addresses with lower amounts, whereas fraudulent transactions tend to involve higher amounts further from the billing address.

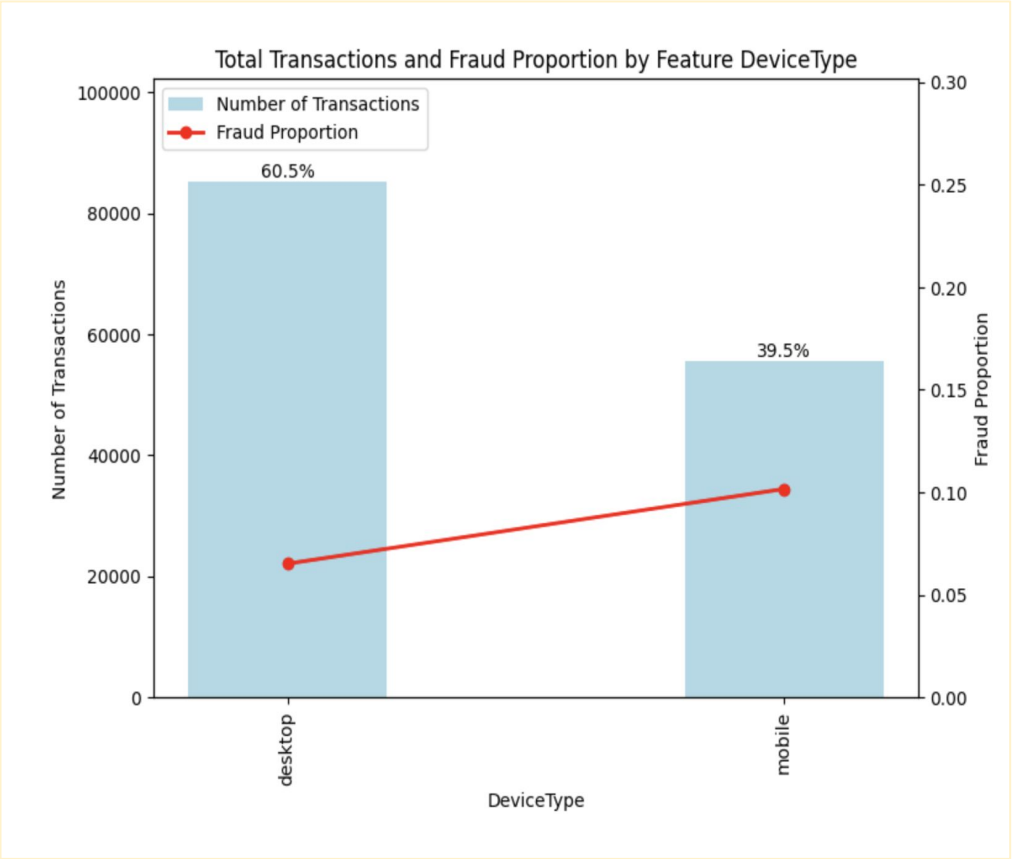
Expected Trends and Patterns [2]

Description:

- The chart compares transactions and frauds across device types: desktop and mobile.
- The bar chart shows transaction distribution, while the red line highlights fraud occurrences per device.

Analysis:

- Desktop accounts for 60.5% of transactions, while mobile covers 39.5%.
- Despite fewer transactions on mobile, fraud rates are higher compared to desktop.



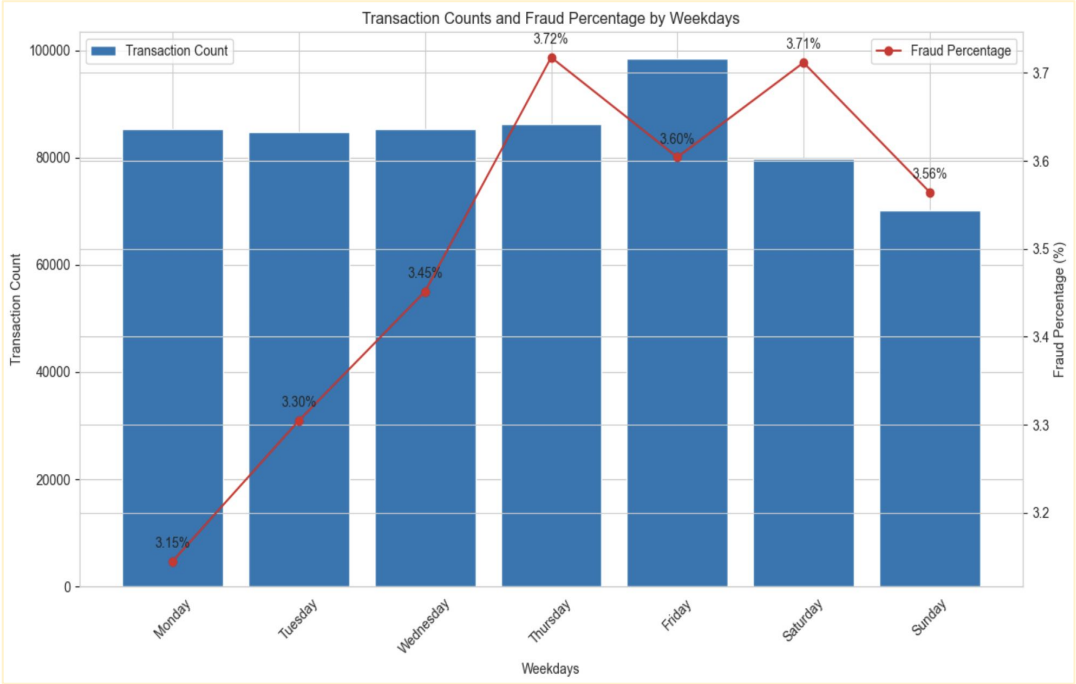
Expected Trends and Patterns [3]

Description:

- The plot displays transaction counts and fraud percentages by weekday.
- The bar chart illustrates the number of transactions for each weekday, while the red line chart represents the fraud percentage in these transactions.

Analysis:

- The number of transactions stays relatively consistent throughout the week, with a slight peak on Friday.
- Fraud percentage is lowest on Monday (3.15%) and rises steadily, peaking on Thursday (3.72%) and maintaining the upward trend over the weekend.



Underlying Trends [1]

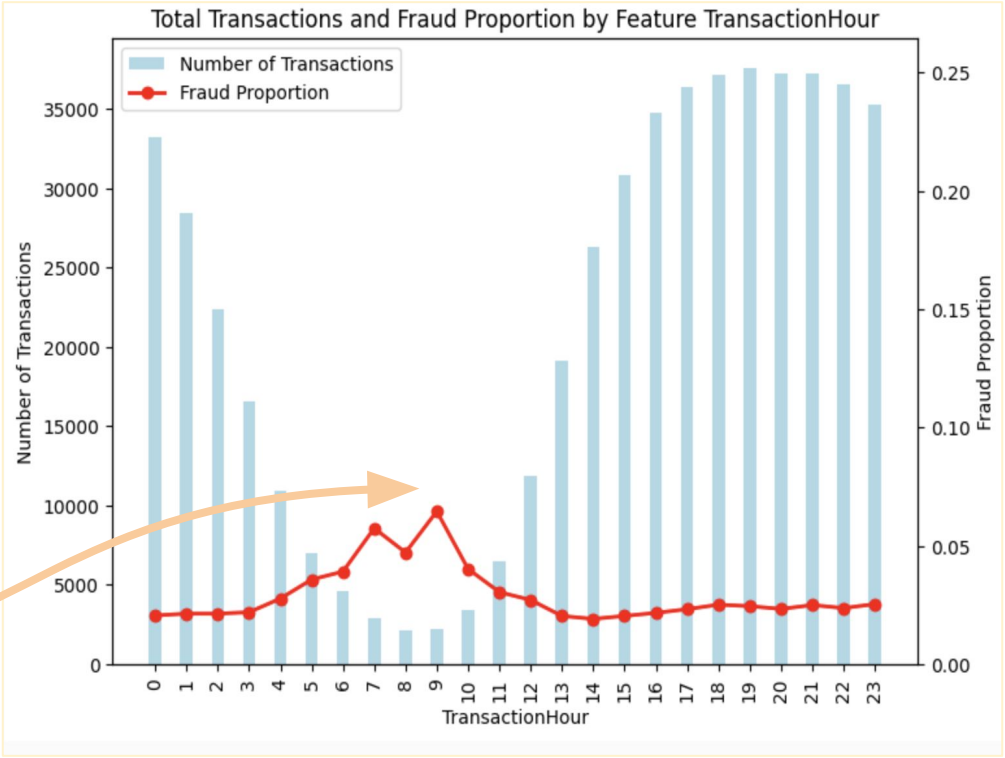
SJSU SAN JOSÉ STATE UNIVERSITY

Description : The graph shows the trend in fraudulent activities with respective to the transaction hour.

Analysis : Transaction timing insights show certain hours show higher transaction volumes, while other hours are notably lower.

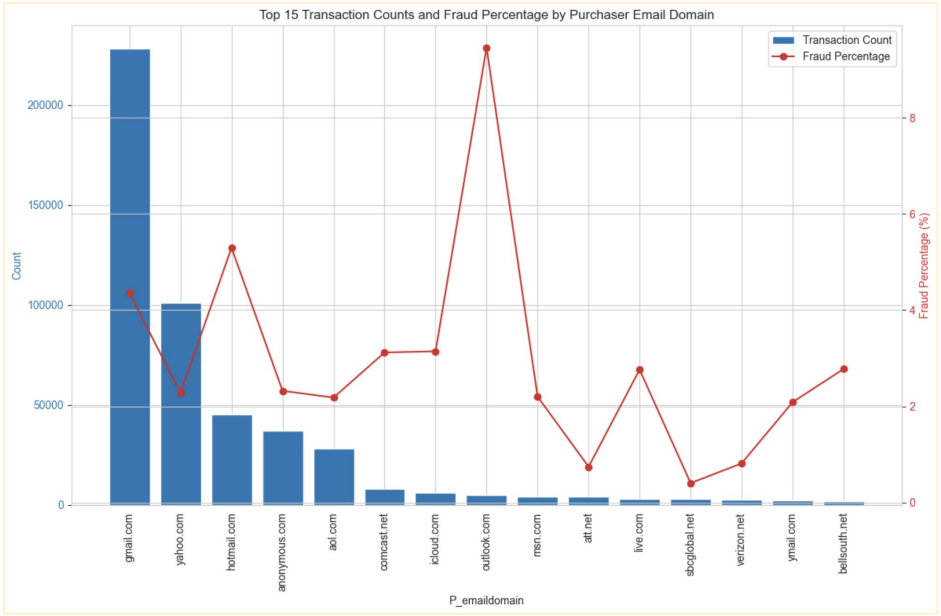
Fraud Patterns:

- Fraudulent activities are more common during hours with fewer transactions.
- This is a notable finding, as it contrasts with the common belief that fraud is more likely to occur during high-transaction periods.
- **Frauds peak during early hours when transaction counts are low.**
- This suggests fraudsters may exploit low-activity periods to evade detection.

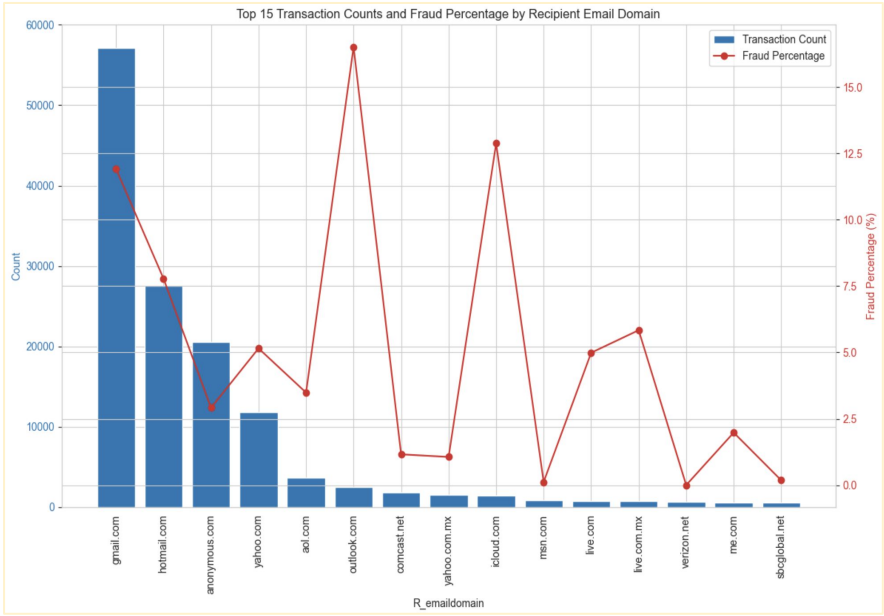


Underlying Trends [2]

SJSU SAN JOSÉ STATE UNIVERSITY



Description : Purchaser transaction counts and fraud percentages for the top 15 email domains, showcasing notable variations in fraud risk.



Description : Recipient transaction counts and fraud percentages for the top 15 email domains.

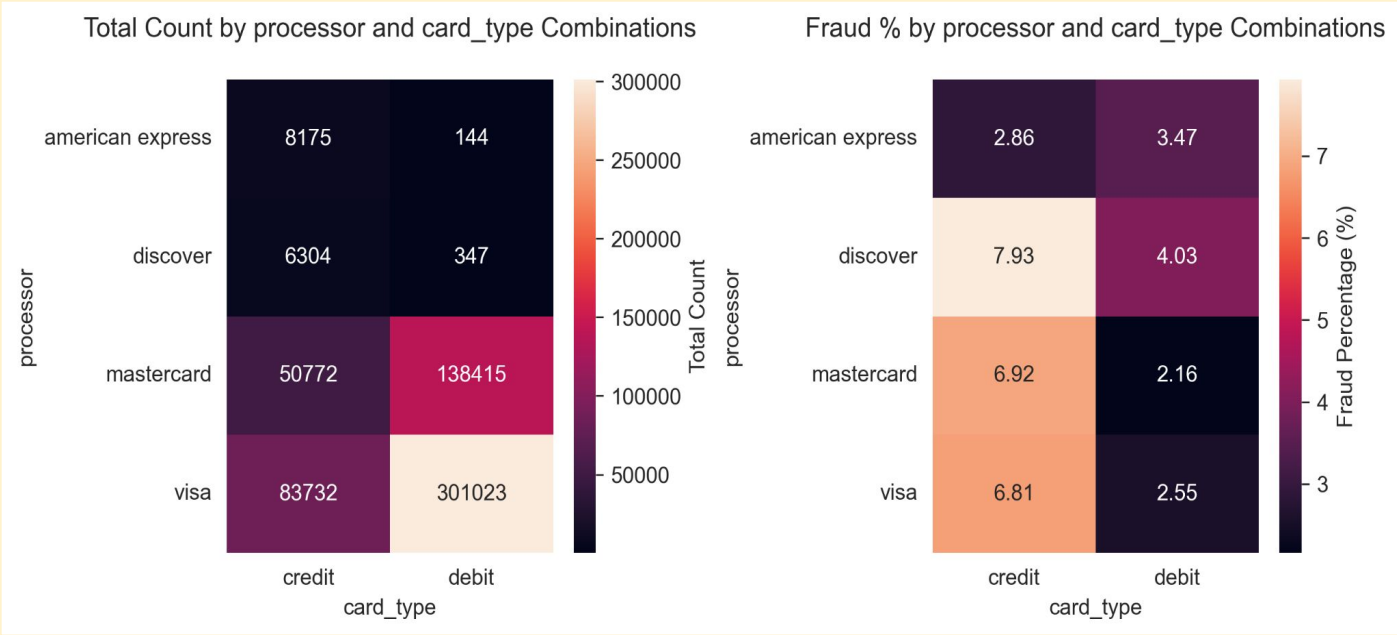
Analysis : High fraud percentages are seen with domains like outlook.com, icloud.com, and gmail.com, suggesting **fraudsters hide behind reputable domains.** These domains are owned by technology giants like Microsoft, Apple, and Google respectively.

Underlying Trends : Card Type vs Fraud [3]

SJSU SAN JOSÉ STATE UNIVERSITY

Description : Heat Map shows number of transactions and fraud percentage for different card companies and card types showing fraud percentages.

Analysis : An unexpected observation is that American Express's debit card has a higher fraud rate than American Express credit card. It contradicts the popular opinion that credit cards are more prone to fraud.



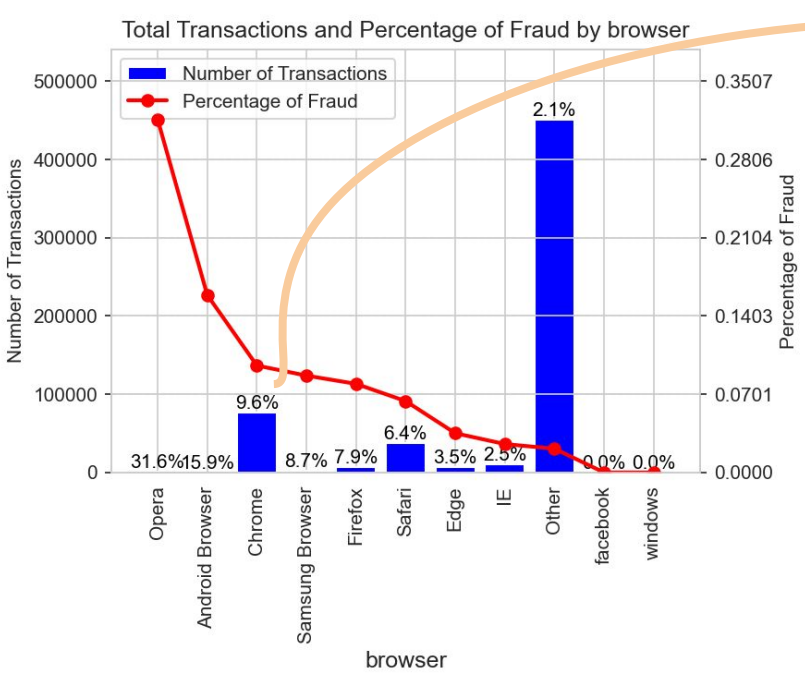
Underlying Trends for Browsers & Chrome [4]

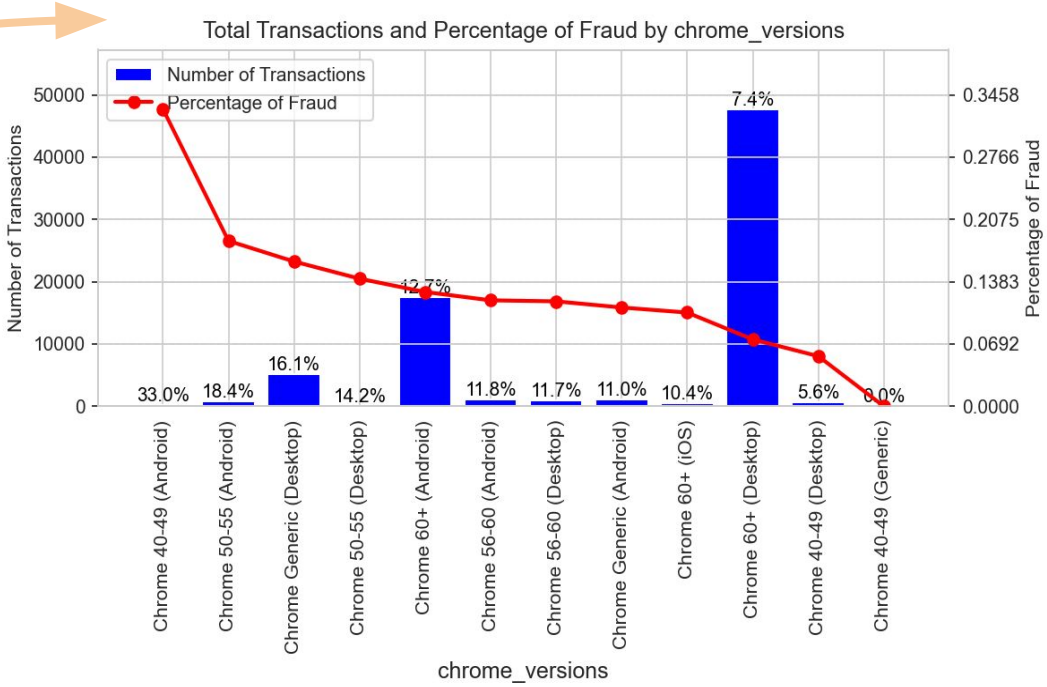
SJSU SAN JOSÉ STATE UNIVERSITY

Description : Figure shows fraud transaction within different browsers and Chrome browser's different versions.

Analysis : Compare to other popular browser, Chrome have highest fraud %. Fraud transactions are higher on mobile Chrome browsers and older desktop versions.

➔ **Newer Chrome versions show lower fraud %, likely due to improved security features.**

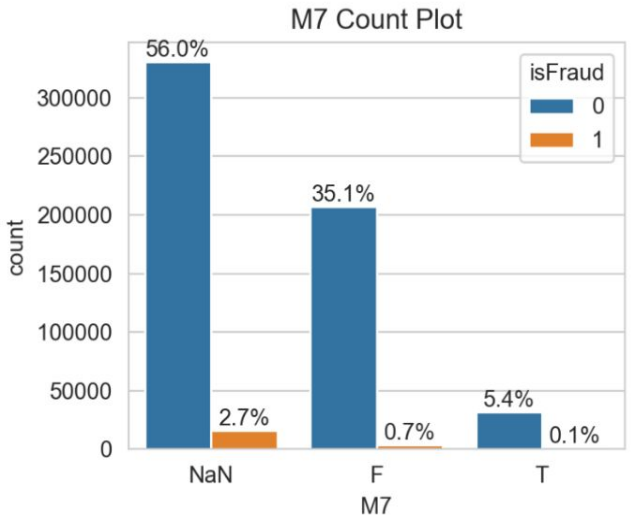


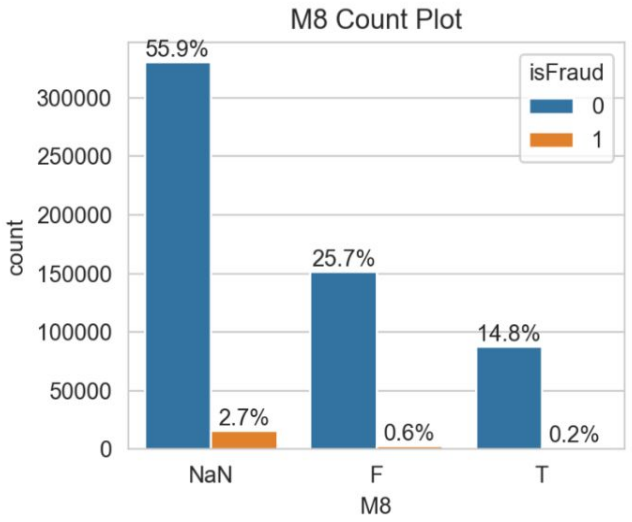


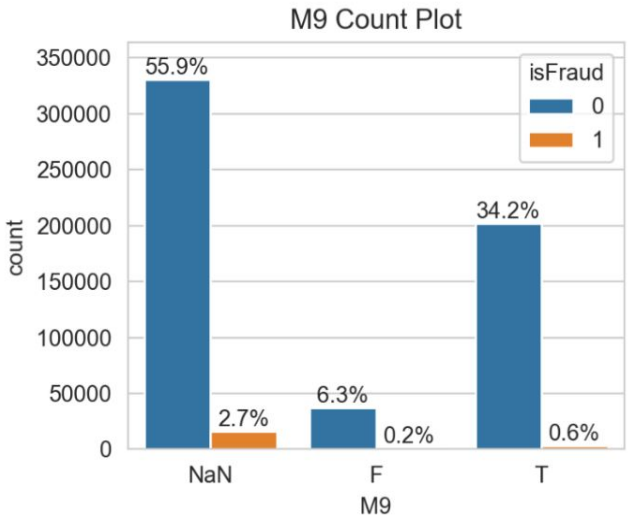
Underlying Trends [5]

Description : Missing Values for the **M-Features (M1-M9),** and the overall fraud transaction contribution which is **2.7% out of total 3.5%**.

Analysis : The M1-M9 features corresponds to the match value such as name on card match or not etc. therefore most of these features should be of the form "T" or "F". But when these variables are not captured during transaction, there is significantly higher chances of fraud. **So it's important to keep all the information up to date with the banks.**

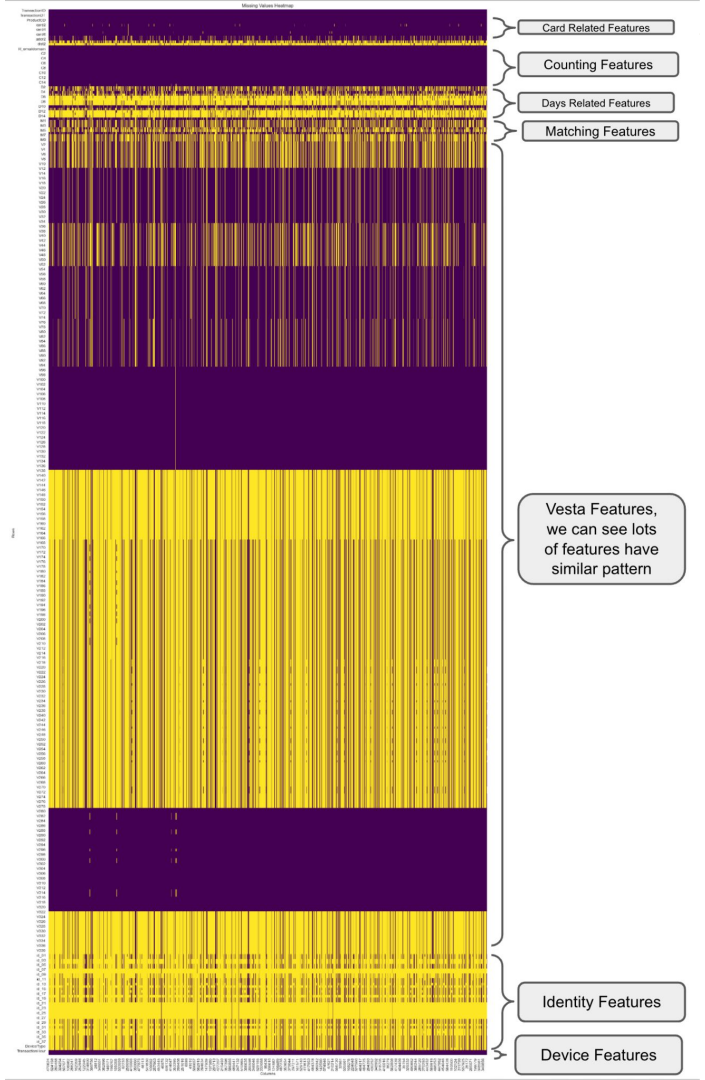






Data Pre-Processing

- Drop columns that have **more than 90 % missing values**. As shown in the figure, there are many columns with lots of missing values.
- Remove outliers using IQR method *.
- Impute missing values -
    a. Using mean for numerical features.
    b. Using mode for categorical features.
- Standardize numerical features using min-max normalization.
- The heatmap plot of the data shows the missing values in data (yellow), and also highlights the recurring pattern of them, which later used for PCA.



* Final models don't use it. We found that removing outliers actually reduced ROC-AUC. This can be explained by the fact that outlier features are indicative of fraudulent behavior.

Feature Engineering [1]

- New features
  - Created correlation-based features to link email domains with fraudulent behavior.
  - Extracted screen resolution (e.g., width x height) into separate features: screen_width and screen_height.
  - Extracted patterns like hour-of-day, day-of-week, and seasonal trends to capture behavioral variations and user habits.

Feature Engineering [2]

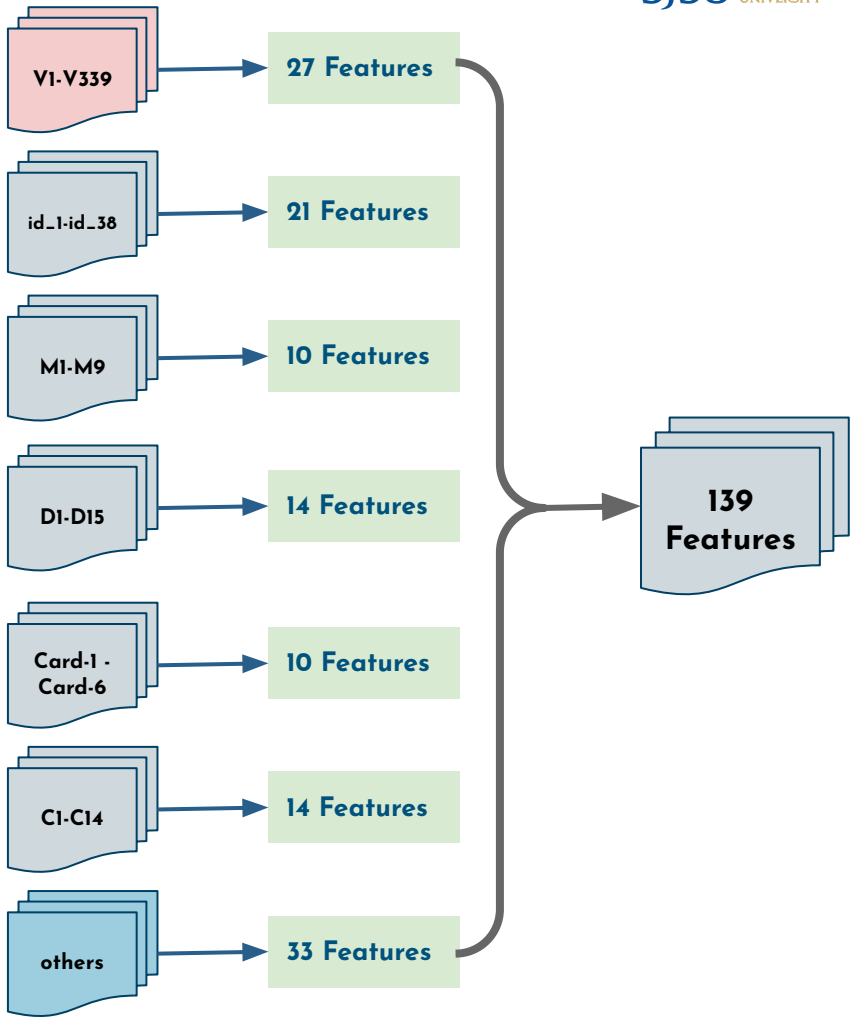
- Feature reduction

  a. Correlated features were grouped, and dimensionality was reduced using PCA. Similarly, email domains (e.g., gmail.com, yahoo.com) were categorized to simplify the feature space.

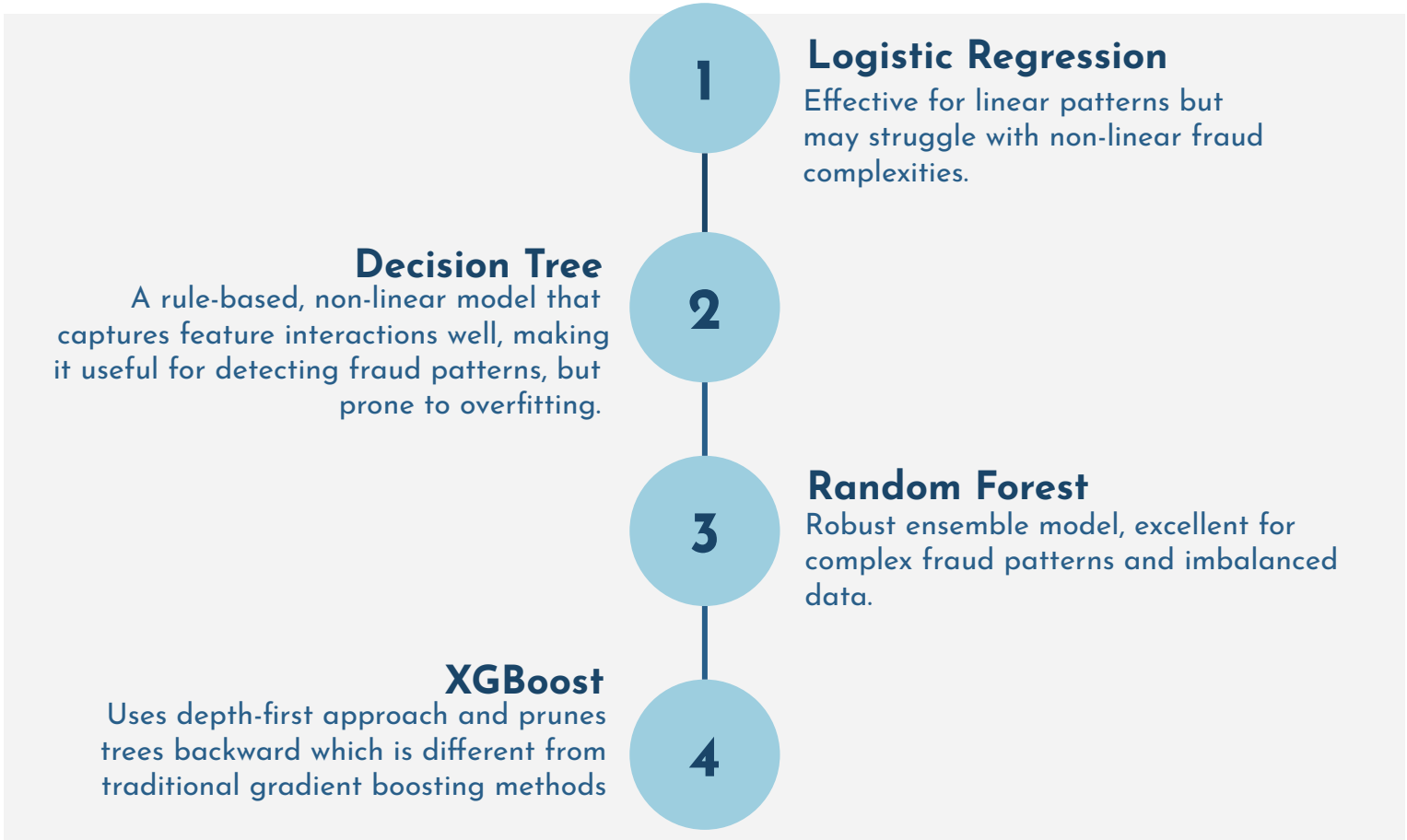
  b. Browsers and their versions were consolidated into broader categories such as Chrome, Firefox, and Safari for better insights.

  c. Operating systems were grouped into broader families, such as Windows, Android, and iOS, for easier analysis..

433 features → 139 features



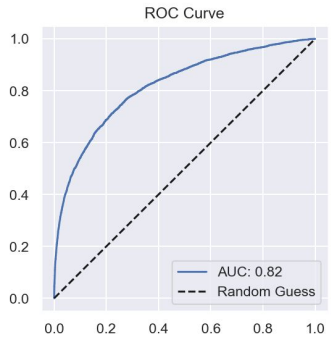
Modelling


1
Logistic Regression
Effective for linear patterns but may struggle with non-linear fraud complexities.
2
Decision Tree
A rule-based, non-linear model that captures feature interactions well, making it useful for detecting fraud patterns, but prone to overfitting.
3
Random Forest
Robust ensemble model, excellent for complex fraud patterns and imbalanced data.
4
XGBoost
Uses depth-first approach and prunes trees backward which is different from traditional gradient boosting methods

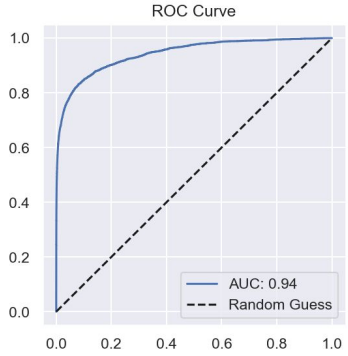
Model Comparison

SJSU SAN JOSÉ STATE UNIVERSITY

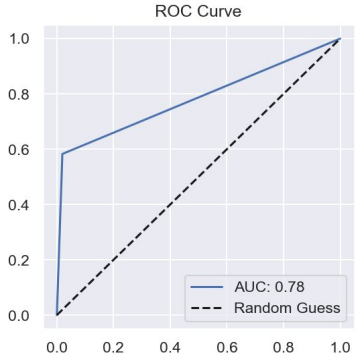
Metrics	Logistic Regression	XGBoost	Decision Tree	Random Forest
Precision	0.6881	0.8963	0.5205	0.9508
Recall	0.0544	0.4892	0.5829	0.4394
F1-Score	0.1009	0.6330	0.5499	0.6010
Accuracy	0.9660	0.9801	0.9666	0.97.96
AUC-ROC	0.8200	0.9430	0.7817	0.9407



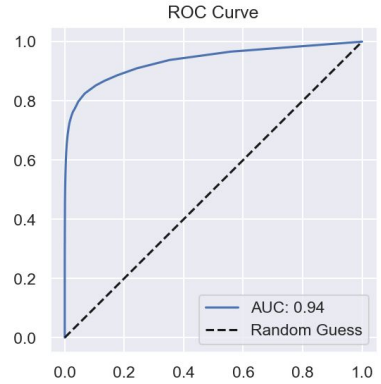
ROC for the Logistic Regression



ROC for the XGBoost



ROC for the Decision Tree



ROC for the Random Forest

Models Analysis : [Random Forest - Feature Importance]

After grid search the top parameters are

'max_depth': 30, '

max_features': 'log2',

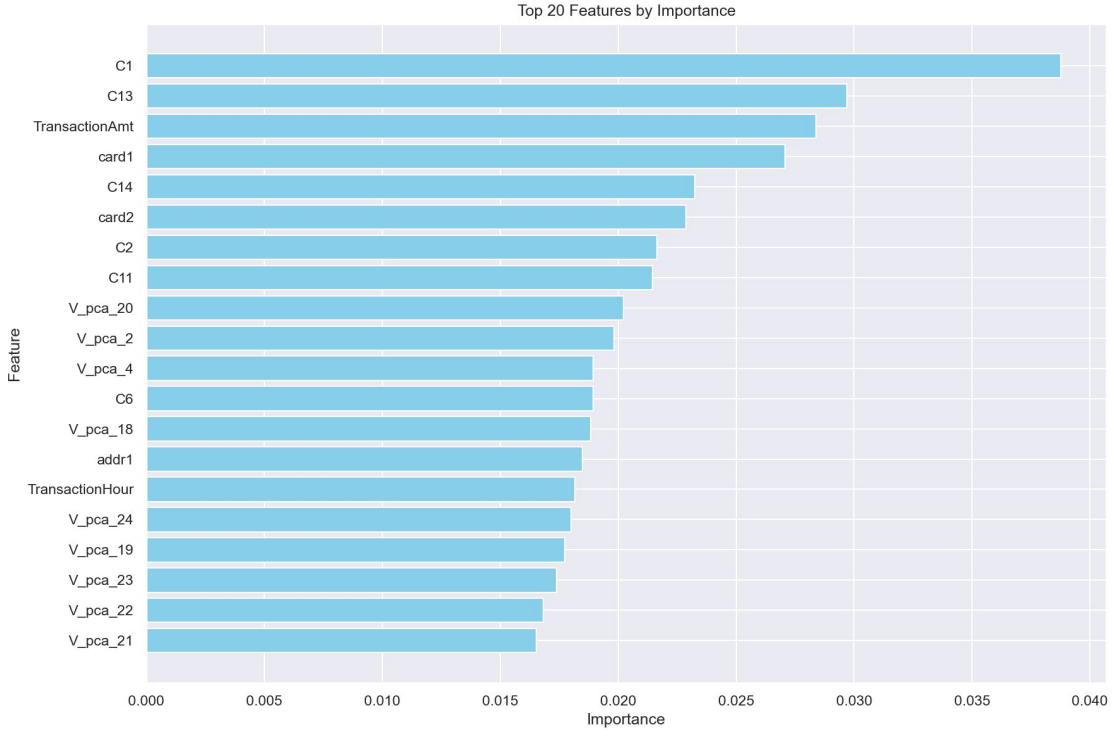
'min_samples_leaf': 1,

'min_samples_split': 2,

'n_estimators': 300

Top 20 selected features used in the random forest algorithm.

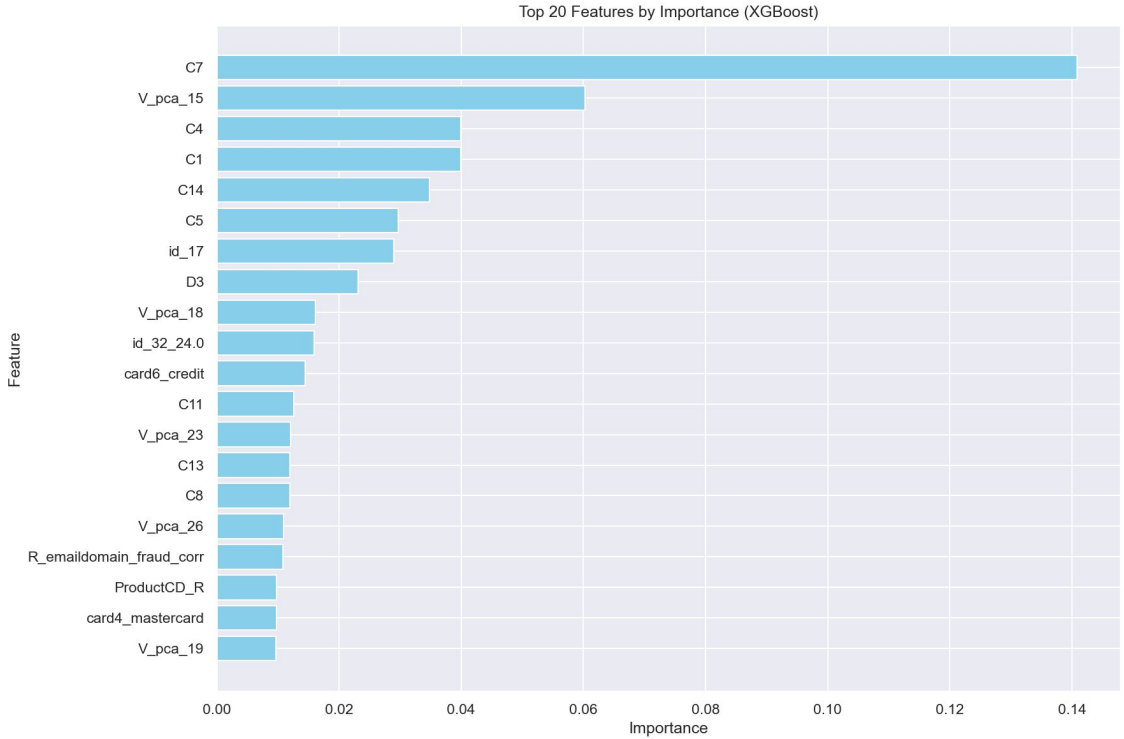
We can see that most of the important features are which we engineered using PCA.



Models Analysis : [XGBoost - Feature Importance]

Top 20 selected features used in the XGBoost algorithm.

We can see that most of the important features are which we engineered using PCA.



Technical Novelty

- **Grouping by Missing Values**: Features were grouped based on missing value patterns to capture similarities.
- **Dimensionality Reduction**: PCA was applied within each group, retaining features that explained 90% of the variance. **339 Vesta** features were reduced to 27.
- **Device Analysis**: Extracted details such as device type, operating system, browser, and screen resolution from raw data.
- **Privacy Constraints**: Due to encoded feature names and data, analysis required relatable interpretation patterns (e.g., D9 represents the hour of the day from normalized Transaction DT revealed only by the histogram pattern.).
- **Model Performance Analysis:** Did further analysis for the transaction where model prediction was wrong, to further understand the data.

Key Findings and Recommendations

SJSU SAN JOSÉ STATE UNIVERSITY

Key Findings

- Fraudulent activity is higher during certain hours, even with fewer transactions occurring at those times.
- Mobile devices are more vulnerable to fraud .
- Popular email domains were more commonly associated with fraudulent activities.
- Older browser versions (e.g., Chrome) show higher fraud rates compared to updated ones.
- In most fraud cases the personal information was not updated in timely manner.
- Contrary to popular belief, debit cards can sometimes be more vulnerable to fraud than credit cards.

Recommendations

- Increased vigilance during high-fraud times.
- Extra caution with mobile transactions, especially regarding email address verification.
- Encourage users to keep browsers and operating system updated to reduce fraud risk.
- Keep the Personal Details like address and Phone number updated with the Issuers.
- Use multi-factor authentication or face ID to access credit card and debit accounts online.

SJSU SAN JOSÉ STATE UNIVERSITY

Community Contribution

1- Guidelines for Fraud Prevention: The project offers data-driven guidelines to help consumers and businesses recognize and avoid fraudulent activity.

3- Encouragement of Safe Practices: Advises consumers to verify email addresses and update browser versions to reduce fraud risks, promoting safe digital habits

5- Increased Consumer Trust: Businesses that actively prevent fraud gain consumer trust, encouraging more support and engagement from the community.

01 02 03 04 05 06

2- Vulnerability Awareness: Highlights time periods and devices (like mobile) that are more prone to fraud, educating the community on higher-risk factors.

4- Enhanced Security: The project provides mechanisms that help protect consumers and small businesses from online fraud and making transactions safer.

6- Financial Resilience for Businesses: Minimizing fraud strengthens businesses financially, allowing them to focus on growth, maintain fair wages, and retain employees.

References

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

Thank You