EDA ANALYSIS REPORT

FITE7410 Financial Fraud Analytics Assignment 1

Date: 7 Mar 2025

Jiang Feiyu [3035770800]

A. Data Cleaning and Preparation Process

*For diagrams, please refer to Appendix 1.

A.1 Missing Value Treatment

Removed 32 features with >50% missing data, including distance variables, time deltas, and identity verification fields. Applied median imputation to 25 numeric variables with partial missingness, including address fields and card-related features. Implemented mode imputation for 17 categorical variables such as payment types and device information.

A.2 Outlier Management

Employed a hybrid approach combining two complementary methods:

- IQR methodology ($M \pm 3 \times IQR/ (2 \times 0.6745)$) for detecting distribution-based anomalies
- Percentile-based truncation (1st-99th percentiles) for addressing extreme tails

Applied the more conservative boundary from both methods, ensuring only truly extreme values were modified (2,504,863 values, representing approximately 25% of total observations). This dual-method approach balanced the need to address influential outliers while preserving legitimate variability in the dataset.

A.3 Additional Preprocessing

Verified complete imputation across variables, with only four identity fields retaining minimal (8,784 values, 2.2%) missingness. Replaced invalid values and conducted consistency checks between related features. Reduced overall dataset dimensions from 101 to 69 variables while maintaining all 100,000 observations

A.4 Data Quality Improvements

Reduced missing values significantly from 3,029,527 to 8,784 (99.7% reduction). Achieved standard deviation reduction of 71.4% on average across treated variables, creating more stable distributions for modeling. Eliminated extreme values that could distort analysis while preserving meaningful variability. Standardized categorical variables through consistent formatting. Implemented cross-field validation to ensure logical consistency between related features.

B. Visualizations of the data

B.1 Univariate Analysis

*For diagrams, please refer to Appendix 2.

Fraud Distribution Analysis: The dataset has 11.3% fraudulent transactions (11,318) and 88.7% legitimate transactions (88,682), showing a moderate imbalance that should be considered during modeling.

Transaction Amount Analysis (Log Scale): The transaction amount distribution is right-skewed with most transactions falling between $25-$100. The log-scale transformation reveals multiple modes, suggesting different transaction amount tiers.

Transaction Amount by Fraud Status: Both fraudulent and legitimate transactions share the same median ($50), but fraudulent transactions show slightly higher variability with more outliers, particularly at higher amounts.

Email Domain Analysis: Google (38,294 transactions) and Microsoft (21,983) dominate as email providers, with Google showing the highest fraud rate among major providers, while telecom providers (AT&T, CenturyLink) show minimal fraud rates.

Product Category Distribution: Product category "key_LY" is most common (44,132 transactions) and has the highest fraud rate (17.3%), while "key_WF" has the highest average transaction amount ($170) but lower fraud rate (5.58%).

Card Information Analysis: Card1 exhibits multiple distinct clusters likely representing different issuing banks or geographic regions; Card2-Card5 shows concentrated distributions around specific values, suggesting they capture categorical card attributes (network types, verification levels, or product tiers).

Device Analysis: Mobile devices account for 38,953 transactions with substantially higher fraud rates (14.5%) than

desktop (9.46% across 58,732 transactions), creating a critical risk signal despite desktop's higher volume; device identifiers predominantly represent mainstream platforms (iOS, Mac, Windows) across both categories.

Fraud Rate by Hour of Day: Morning hours (7-10am) show dramatically elevated fraud rates (29-35%), far exceeding other times of day, creating a striking temporal pattern for fraud activity.

C Feature Distribution: Exhibit highly right-skewed distributions with low medians (0-1) but high means, suggesting binary or count features capturing rare events. Heavy concentration at low values with occasional extreme outliers indicates risk flag or trigger count variables.

D Feature Distribution: Dominated by zero values (most medians = 0) with substantial right tails and high maximums, characteristic of time-based deltas or cumulative metrics. D2 shows distinct pattern with higher median (36), potentially representing a more commonly triggered temporal measurement.

B.2 Bi-/Multi-variate Analysis

*For diagrams, please refer to Appendix 3.

B.2.1 Fraud Indicator Correlations

Among the C and D feature families, we identified several variables with statistically significant correlations to fraudulent transactions:

- Positive fraud indicators: The highest positive correlations with fraud were observed in C2 (0.047), C1 (0.040), and C12 (0.035), suggesting these card verification-related variables are reliable indicators of potentially fraudulent activity. (refer to Table 1 in Appendix 3)
- Negative fraud indicators: The strongest negative correlations were found in D1 (-0.051) and C3 (-0.029), indicating these variables are more commonly associated with legitimate transactions. C13 (0.015), while technically positive, showed the lowest correlation with fraud. (refer to Table 2 in Appendix 3)

B.2.2 Bivariate Relationship Patterns

When visualized through boxplots and scatter plots, the data revealed three key patterns:

Distributional differences: Clear separation in the distribution of key features between fraudulent and legitimate transactions, particularly for C2 and C1 which show consistently higher values in fraudulent cases.

Feature redundancy: Nearly perfect linear relationships between certain feature pairs (C6-C4, C12-C7), indicating significant collinearity issues that should be addressed prior to modeling to prevent computational inefficiency and potential model instability.

Combined predictive power: While individual features show moderate but significant correlations with fraud status, their limited individual strength suggests that a multivariate approach leveraging feature interactions would yield substantially better results than univariate methods. (refer to Table 3 in Appendix 3)

B.2.3 Implications for Modeling

These bivariate and multivariate findings have important implications for our modeling approach:

Features C2, C1, and C12 should be prioritized in model development as primary fraud indicators. Highly correlated feature pairs should be evaluated for dimension reduction to improve model efficiency. The modest individual correlations but clear distributional differences suggest ensemble methods or models capable of capturing non-linear

relationships would be most effective.

C. Feature Engineering

*For diagrams, please refer to Appendix 4.

27 features are engineered using multiple techniques to transform raw transaction data into powerful fraud signals. The implementation successfully enhanced the dataset's predictive capacity through:

- Temporal Pattern Extraction: Derived time components revealing high-risk morning hours (7-10am) with 29-35% fraud rates, significantly above average.
- Statistical Transformations: Applied log transformations to transaction amounts and created percentile rankings to normalize distributions and identify outliers.
- Risk-Based Encoding: Converted categorical variables (email providers, device types) into numerical risk scores, revealing Google emails and mobile devices (14.5% fraud rate vs desktop's 9.46%) as high-risk indicators.
- Threshold Flagging: Implemented binary flags for high-risk values in features with strongest fraud correlations (C2: 0.047, C1: 0.040).
- Aggregation & Segmentation: Calculated group-level fraud rates for card issuers and product categories, identifying 'key_LY' products with 17.3% fraud.
- Composite Scoring: Created standardized risk scores combining multiple indicators into holistic risk assessments.
- Interaction Features: Developed combinations of high-risk factors to capture complex fraud patterns, particularly focusing on risky time-device combinations.

The features significantly improved discriminative power as confirmed by visualizations showing clear separation between fraud and legitimate transaction distributions in the engineered features.

D. Discussion of Key Findings

The EDA revealed several significant fraud patterns and generated actionable hypotheses. Temporal analysis identified morning hours (7-10am) as high-risk periods with fraud rates reaching 35%, suggesting fraudsters target times when monitoring may be reduced. Device usage patterns showed mobile transactions experiencing 53% higher fraud rates than desktop transactions (14.5% vs 9.46%), indicating potential security vulnerabilities in mobile platforms. Email domain analysis identified Google as a high-risk provider, while product category "key_LY" exhibited a 17.3% fraud rate, significantly above average. Card verification features (C2, C1, C12) demonstrated the strongest correlation with fraud (0.047, 0.040, 0.035), while feature pairs showed significant collinearity, suggesting dimension reduction opportunities. These findings support our hypothesis that fraud detection requires a multi-dimensional approach combining temporal, behavioral, and verification signals rather than relying on individual indicators. The distinct distributional differences between legitimate and fraudulent transactions across engineered features confirm our hypothesis that transformed variables significantly enhance discriminative power.

Appendix

Appendix 1. *Data Cleaning and Preparation Process*

Data Cleaning Summary Report

1. Missing Values Treatment

- Removed 32 columns with >50% missing values

- Applied median imputation to 25 numeric columns

- Applied mode imputation to 17 categorical columns

2. Outlier Treatment

- Applied combined IQR and percentile-based approach

- Treated outliers in numeric variables

- Total outliers modified: 2504863

3. Dataset Dimensions

- Original: 100000 rows, 101 columns

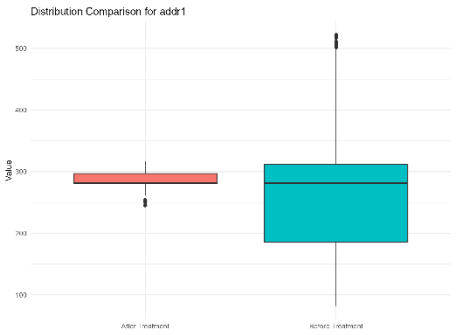
- Final: 100000 rows, 69 columns

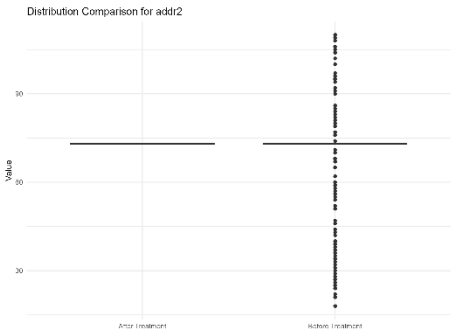
- Features removed: 32

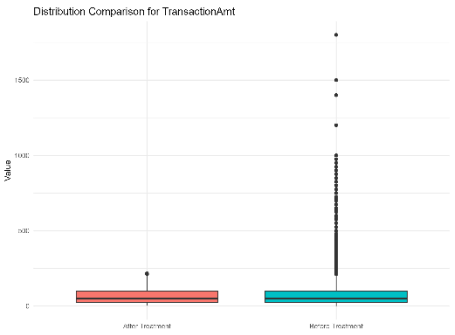
4. Data Quality Improvements

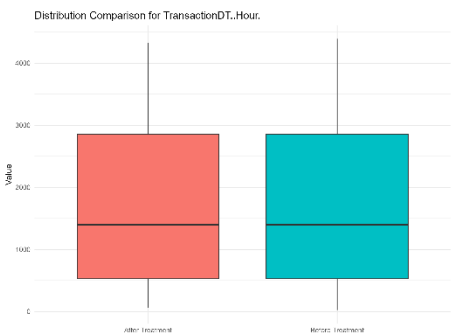
- Missing values reduced from 3029527 to 8784

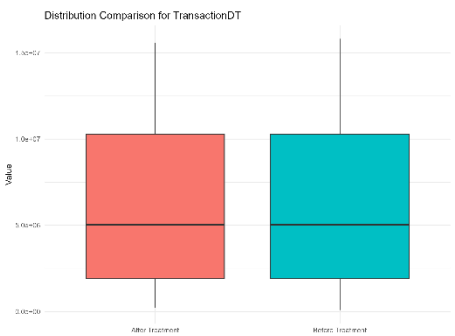
- Standard deviation reduction (average across treated variables): 71.4 %

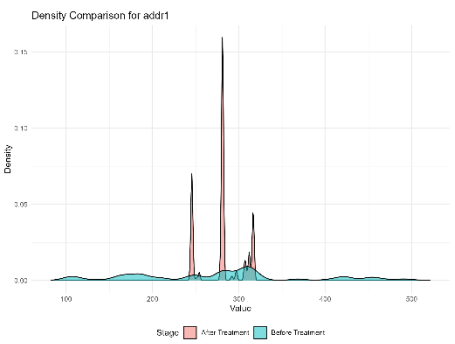


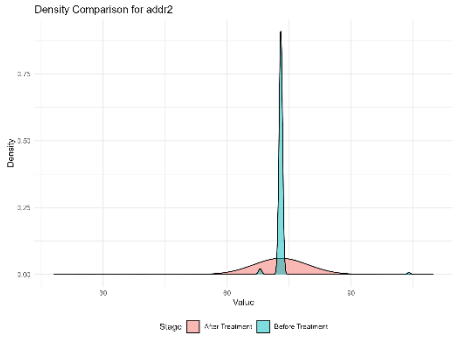


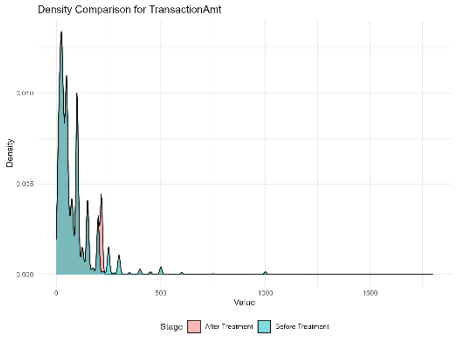


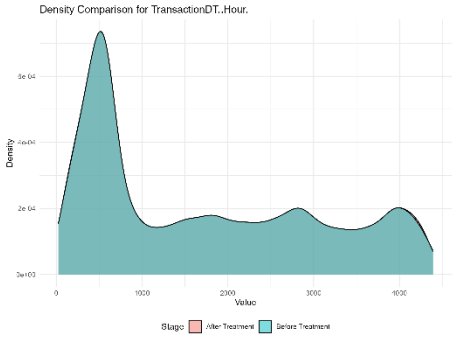


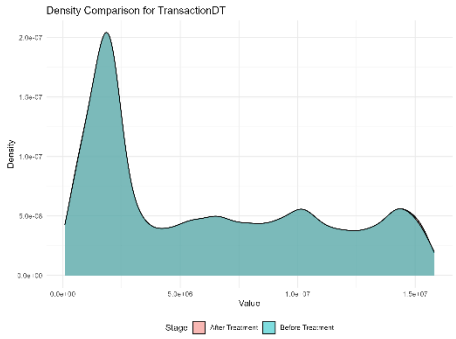


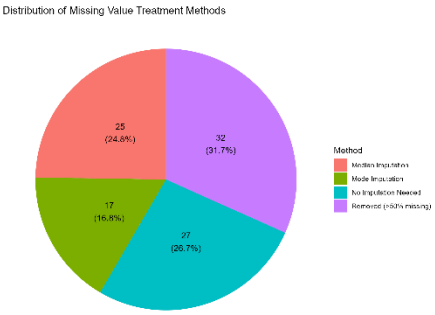


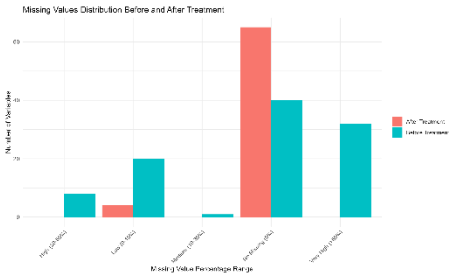

Density Comparison for addr2
Density
Value
Stage
After Treatment
Before Treatment

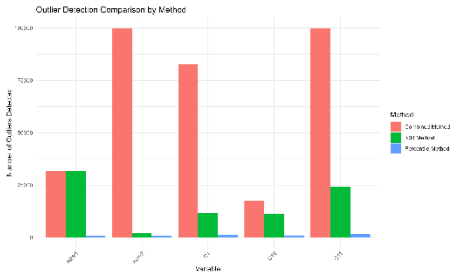

Density Comparison for TransactionAmt
Density
Value
Stage
After Treatment
Before Treatment


Density Comparison for TransactionDT..Hour.
Density
Value
Stage
After Treatment
Before Treatment

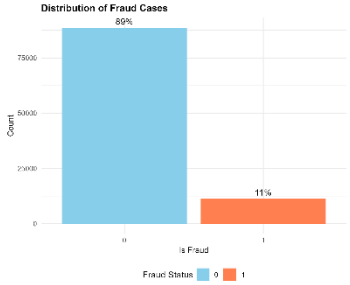

Density Comparison for TransactionDT
Density
Value
Stage
After Treatment
Before Treatment

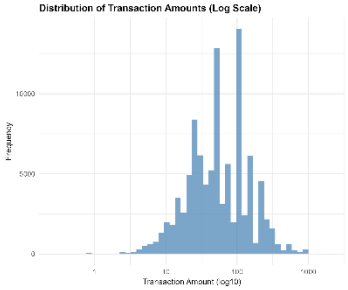

Distribution of Missing Value Treatment Methods
25
(24.8%)
32
(31.7%)
17
(16.8%)
27
(26.7%)
Method
Median Imputation
Mode Imputation
No Imputation Needed
Removed (>50% missing)

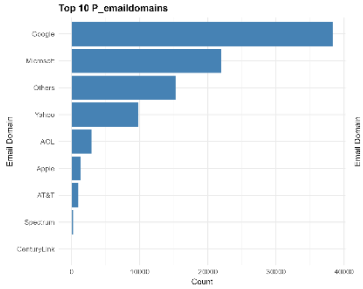

Missing Values Distribution Before and After Treatment
Missing Value Percentage Range

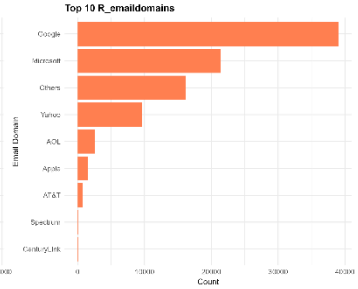

Outlier Detection Comparison by Method
Variable
Method

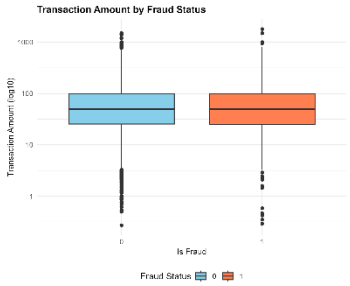
Appendix 2. *Univariate Analysis*

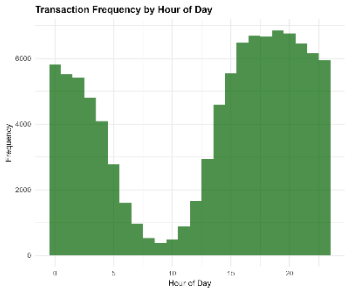


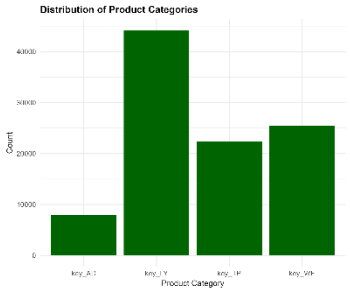


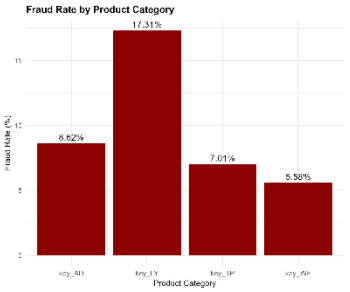


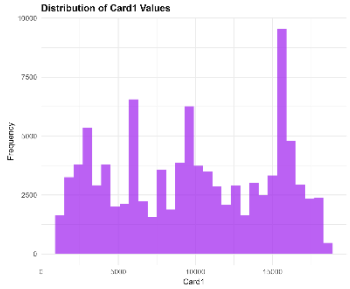


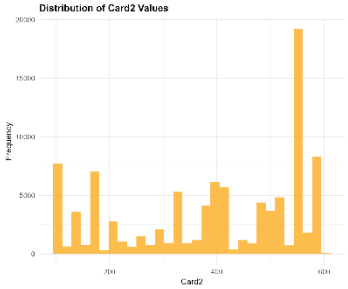


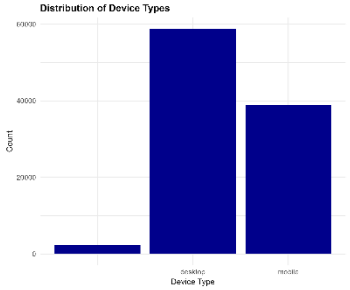


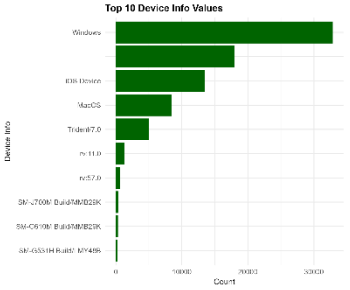


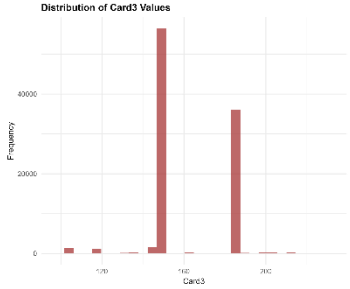


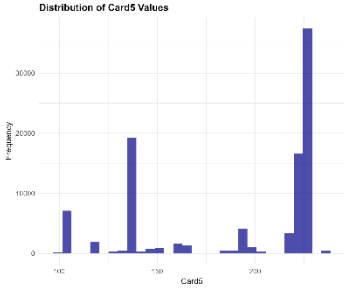


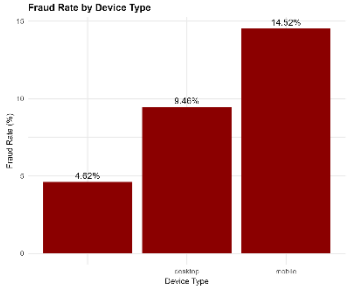


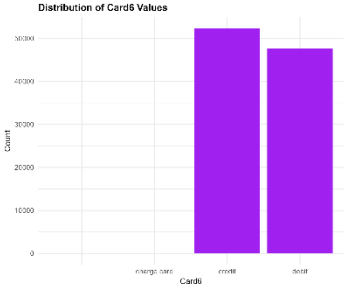


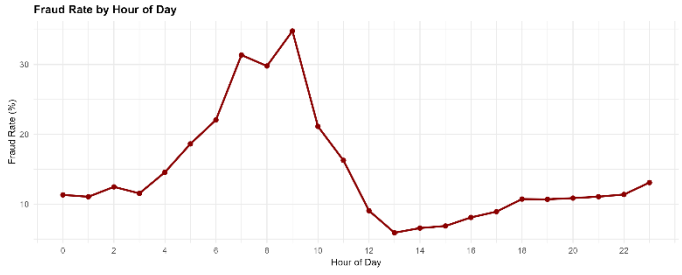


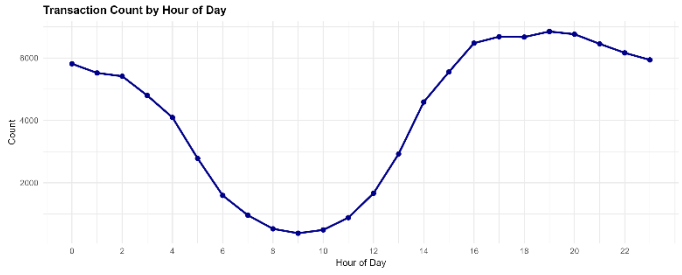


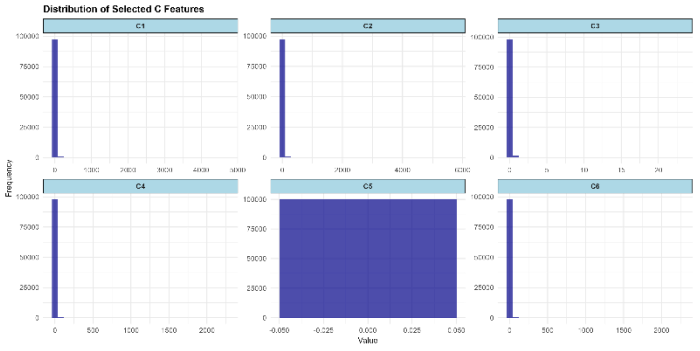


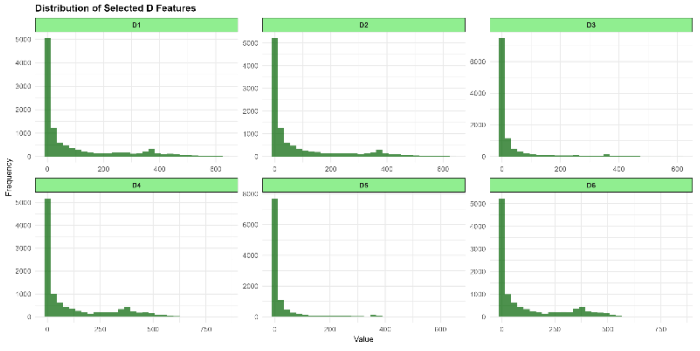












Appendix 3. ***Bi-/Multi-variate Analysis***

- C2	C1	C12	C11	C8
0.04654697	0.03963151	0.03469159	0.03362412	0.03186323

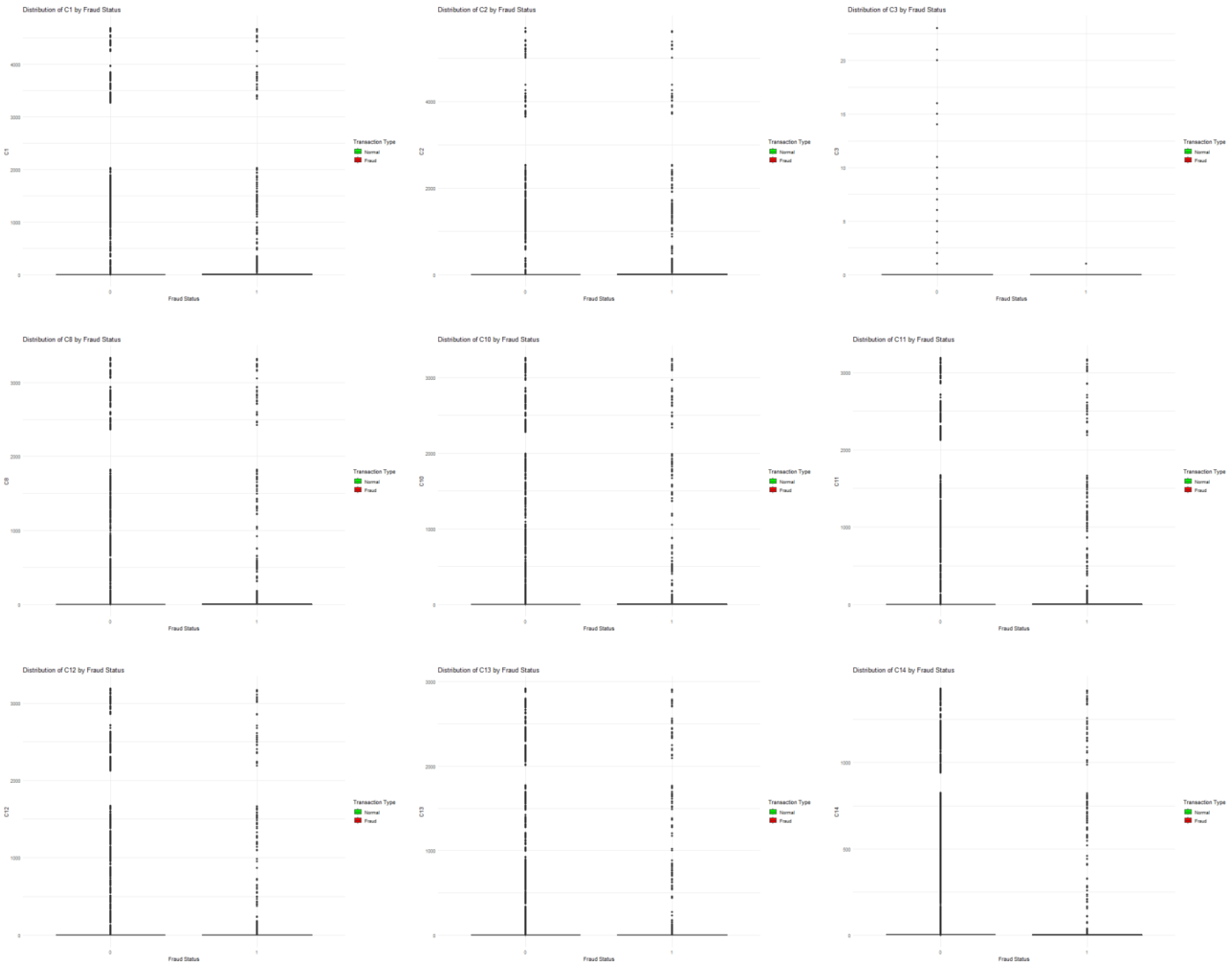
Table 1: Top positive correlations with fraud

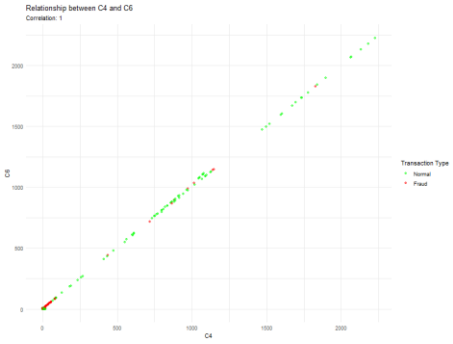
C10	C14	C13	C3	D1
0.02523586	0.01628589	0.01490276	-0.02939635	-0.05144087

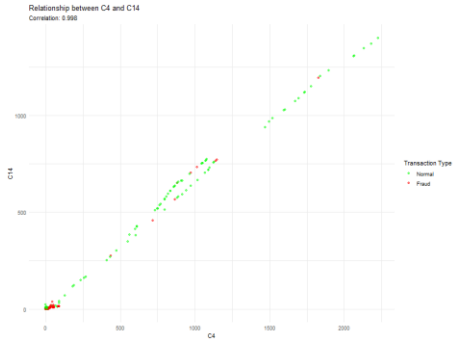
Table 2: Top negative correlations with fraud

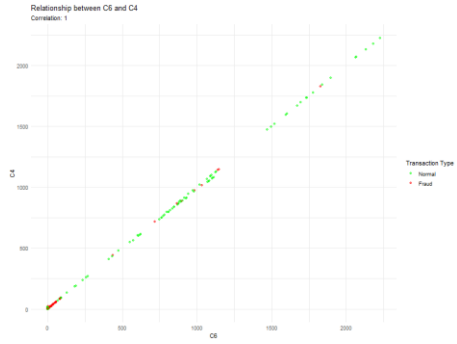
Feature1	Feature2	Correlation
C6	C4	0.9999373
C4	C6	0.9999373
C12	C7	0.9993554
C7	C12	0.9993554
C14	C6	0.9985054

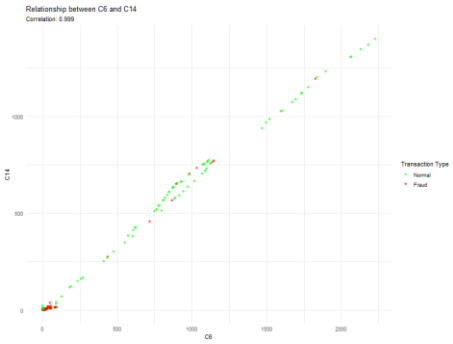
Table 3: Top correlated feature pairs

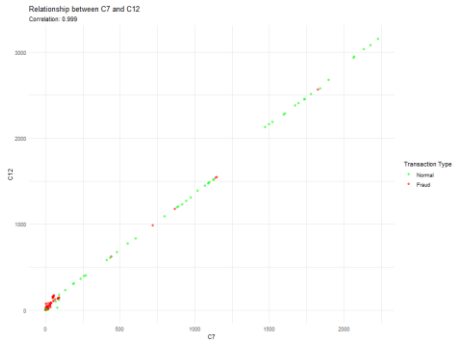


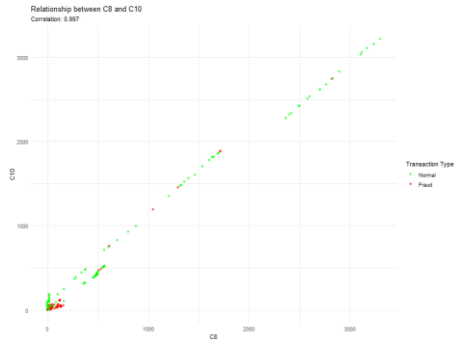

Relationship between C4 and C6
Correlation: 1
C6
C4
Transaction Type
Normal
Fraud

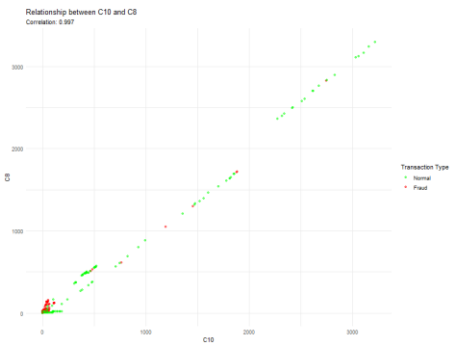

Relationship between C4 and C14
Correlation: 0.998
C14
C4
Transaction Type
Normal
Fraud

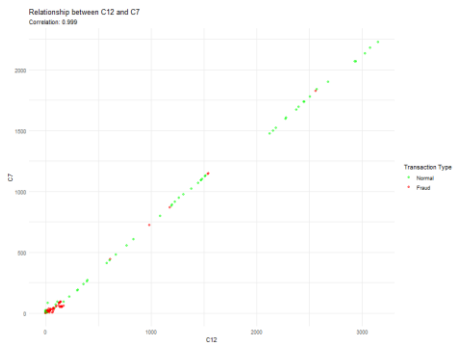

Relationship between C6 and C4
Correlation: 1
C4
C6
Transaction Type
Normal
Fraud

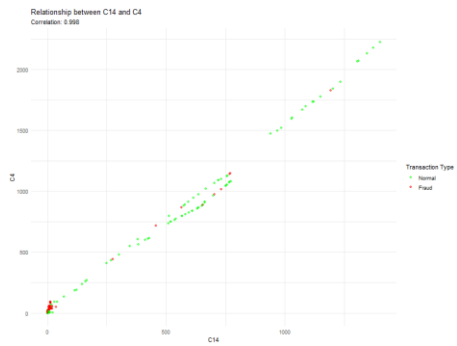

Relationship between C6 and C14
Correlation: 0.999
C14
C6
Transaction Type
Normal
Fraud

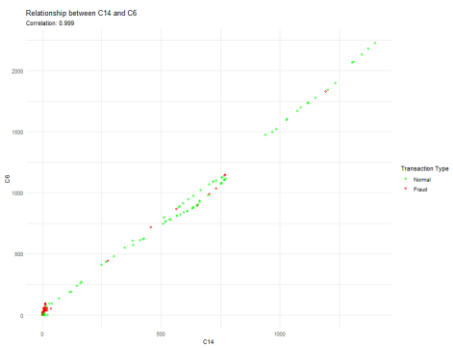

Relationship between C7 and C12
Correlation: 0.999
C12
C7
Transaction Type
Normal
Fraud

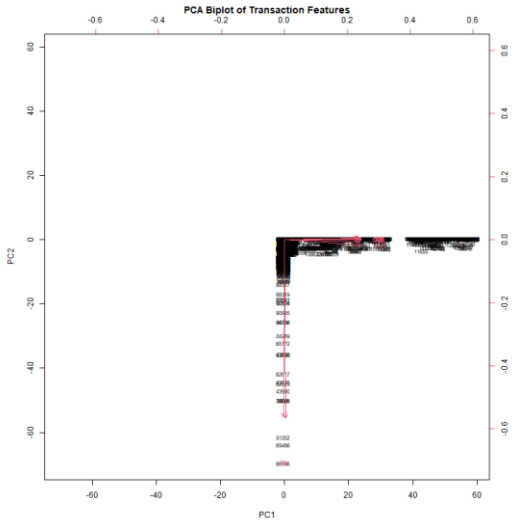

Relationship between C8 and C10
Correlation: 0.997
C10
C8
Transaction Type
Normal
Fraud

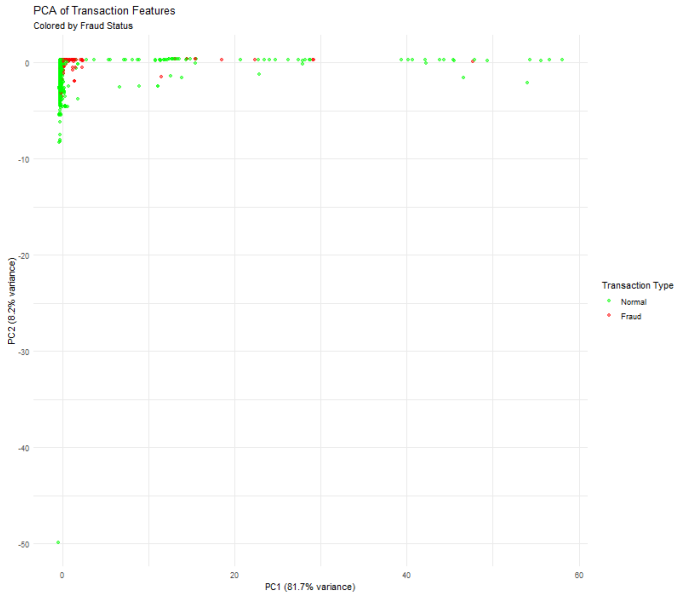

Relationship between C10 and C8
Correlation: 0.997
C8
C10
Transaction Type
Normal
Fraud


Relationship between C12 and C7
Correlation: 0.999
C7
C12
Transaction Type
Normal
Fraud

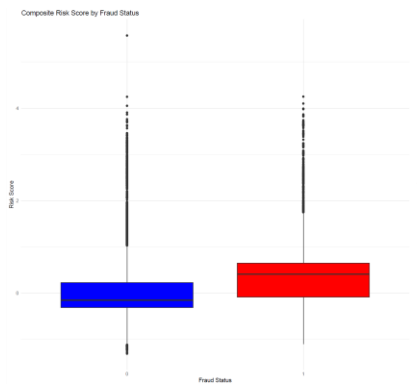

Relationship between C14 and C4
Correlation: 0.998
C4
C14
Transaction Type
Normal
Fraud

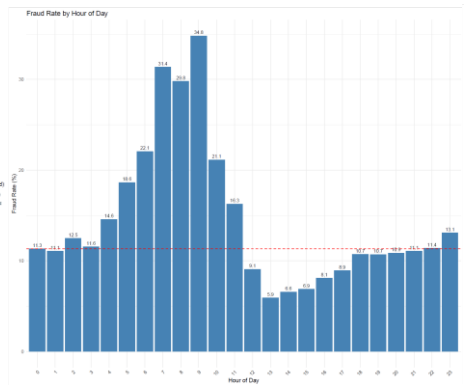

Relationship between C14 and C6
Correlation: 0.999
C6
C14
Transaction Type
Normal
Fraud

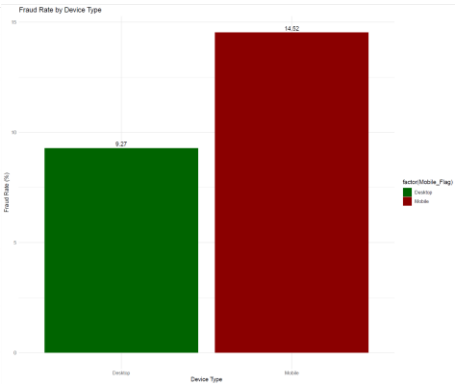


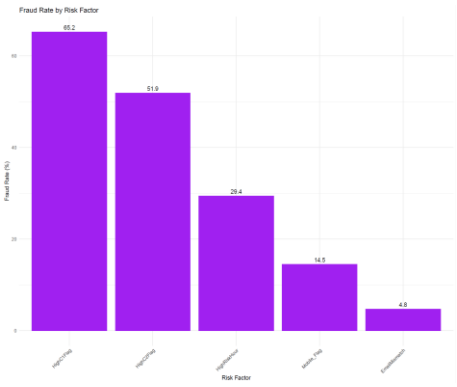


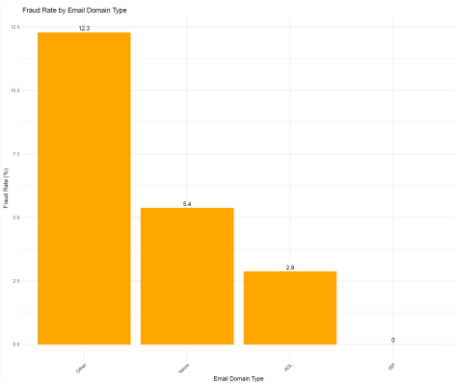
Appendix 4. *Feature Engineering*











Feature Engineer Summary

Time-based Features:

- Transaction_Hour: Hour of the day (0-23)
- Transaction_DayOfWeek: Day of week (1-7, Monday-Sunday)
- Transaction_Weekend: Binary weekend indicator
- DayPeriod: Categorized period of day (Night, Morning, Afternoon, Evening)
- HighRiskHour: Flag for high-risk morning hours (7-10am) with fraud rates 29-35%

Amount-based Features:

- LogAmount: Log-transformed transaction amount to handle skewness
- AmountPercentile: Percentile rank of transaction amount
- AmountDeviation: Deviation from product category median
- AmountDeviationRatio: Ratio to product category median

Email Domain Features:

- P_emaildomain_type: Categorized email provider
- DomainFraudRate: Fraud rate associated with email domain

Device Features:

- Mobile_Flag: Flag for mobile transactions (higher fraud risk)
- DeviceFraudRate: Fraud rate associated with specific device

Transaction Behavior Features:

- HighC2Flag/HighC1Flag: Flags for high-risk C values (top 5%)
- CRiskScore: Combined risk score based on C2, C1, C12 correlations

Verification Mismatch Features:

- EmailMismatch: Flag for mismatch between purchaser and recipient email domains
- CardMismatchFlag: Aggregate of card-related verification flags

Aggregated Behavior Features:

- Card1FraudRate: Fraud rate associated with card issuer
- ProductFraudRate: Fraud rate associated with product category

Combined Risk Score:

- CompositeRiskScore: Standardized composite of multiple risk indicators

Interaction Features:

- HighRiskHour_Mobile: High-risk hour on mobile device
- HighAmount_HighRiskHour: Large transaction during high-risk hours

HighRisk_EmailMismatch: Email mismatch for transactions with high composite risk