E-commerce Fraud Detection

By Valentina Sanchez

Introduction to e-commerce fraud and its impact

ISSUE

With the surge in online shopping, especially after the COVID-19 pandemic, e-commerce has thrived. However, this growth also attracted fraudulent activities.

CONCERN

E-commerce fraud is a significant concern for businesses and banks. As reported in a PYMNTS.com article, '70% of card-related fraud occurs in a card-not-present (CNP) scenario.' CNP fraud is expected to reach $49 billion globally by 2030."

Who might care?

BANKS

BUINESSES

CONSUMERS

and many more...

FACTORS

Contributing Fraud

Transaction Patterns: Anomalies in transaction patterns, such as unusual purchase frequency or payment methods, can signal potential fraud.

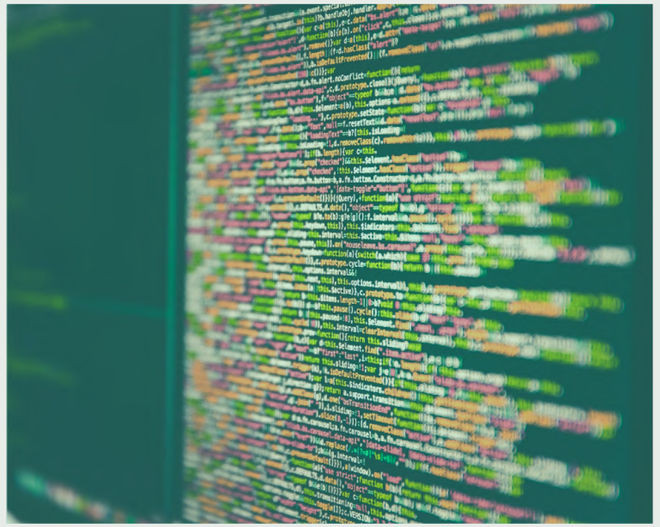
Transaction Times: The timing of transactions can also be a critical factor; for instance, late-night or unusual transaction times may indicate fraud attempts.

Transaction Amounts: Unusual or exceptionally large transaction amounts can be indicators of fraudulent activity.

DATA

Sources

Researches from the IEE Computational Intelligence Society (IEE CIS) have partnered with Vesta Corporation, a payment service company, to seek the best solutions for fraud prevention industry.
Presenting a challenging large-scale dataset with a wide range of anonymous features from real-life scenarios.



Data Information

IDENTITY TABLE (144233, 41)

TRANSACTION TABLE (590540, 394)

Left join on the 'TransactionID' to avoid excluding any transaction-related information resulting in a shape of (590540, 434)

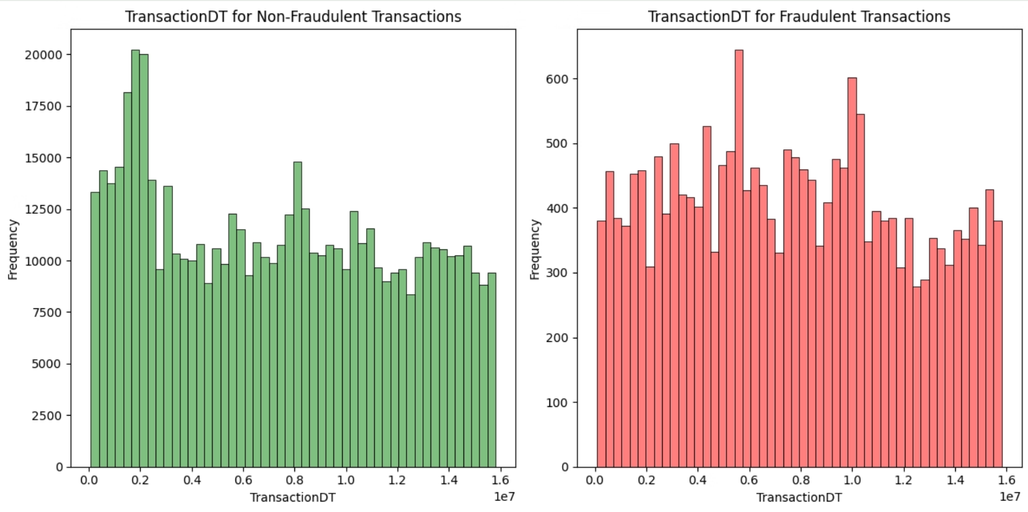
Our dataset is sourced from Vesta's fraud protection system and digital security partners. It contains valuable variables related to identity information, network connections (IP, ISP, Proxy, etc.), and digital signatures (UA/browser/os/version, etc.) associated with transactions.

Please note that for privacy and contractual reasons, field names are masked, and a pairwise dictionary were not provided.
https://www.kaggle.com/competitions/ieee-fraud-detection/overview

Historical Data

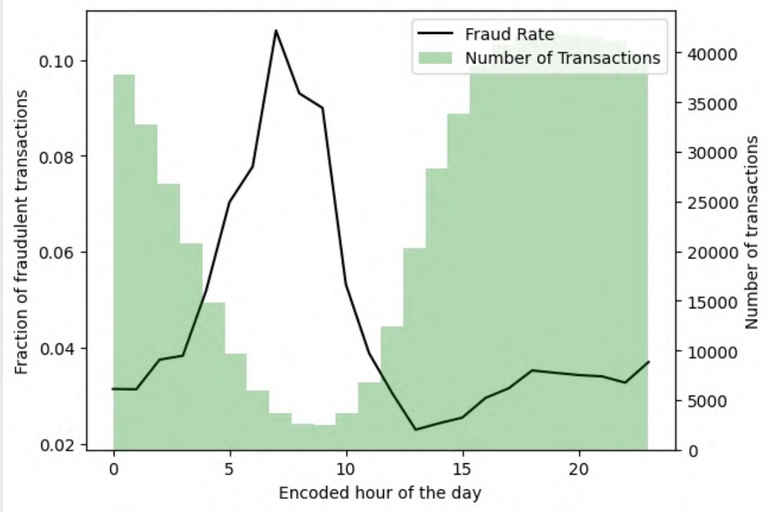
SOME FEATURE ENGINEERING

OBSERVING TRANSACTIONDT



Data timestamps are represented as timedeltas from a reference datetime, rather than actual timestamps.

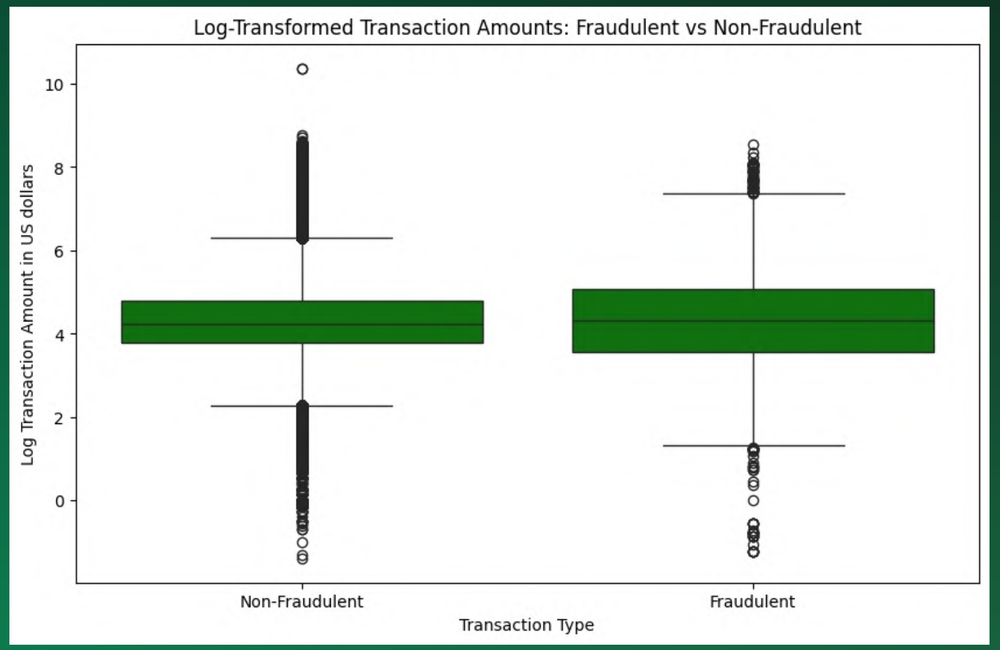
FEATURE ENGINEERING TRANSACTIONDT



The 'hour_of_the_day' feature is created from TransactionDT timestamps, providing the encoded hour. The analysis helps in understanding when fraud is most likely to occur during the day.

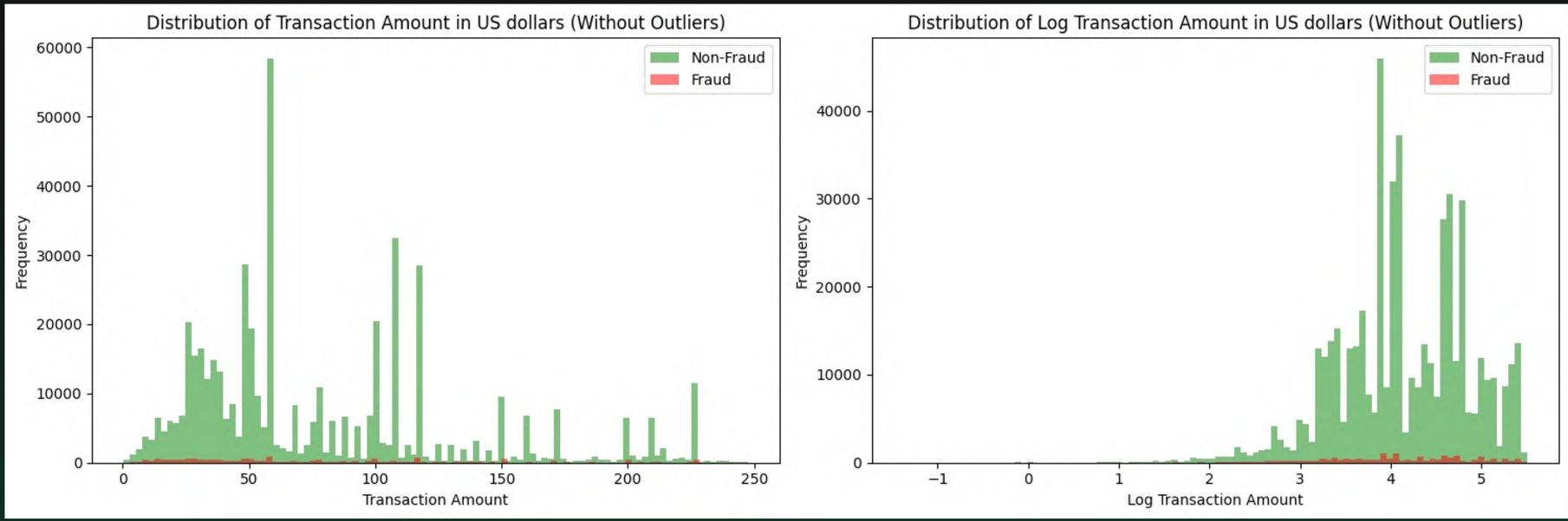
Target Variable isFraud

HANDLING UNBALANCED DATA AND IMPUTATION STRATEGY



- Unbalanced datasets are common in fraud detection scenarios.
- In our dataset, many features had substantial missing data (around 80%).
- **Decision**: Instead of feature removal, a strategic approach was adopted.
- **Categorical Features**: Imputed using the most frequent values.
- **Numerical Features**: Imputed using the median values.
- This approach preserves valuable information while addressing missing data challenges.
- Ensures a more robust and comprehensive analysis for fraud detection.

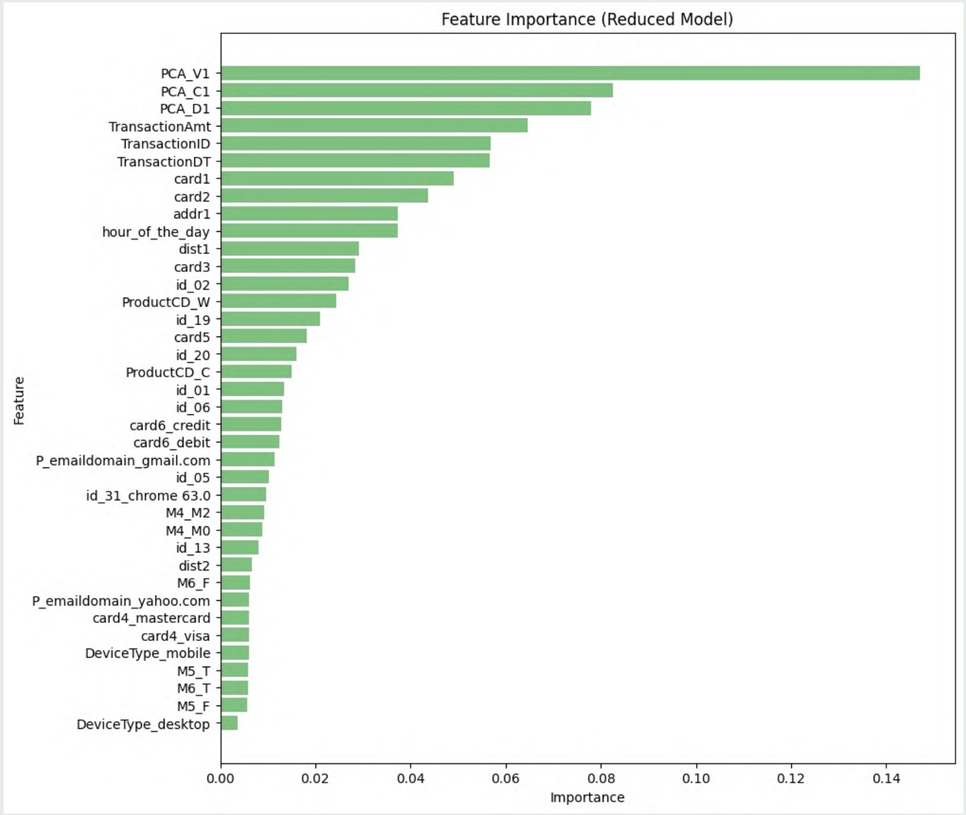
TransactionAmt



It is evident that fraudulent transactions can occur at any scale, underscoring the need for vigilance across all transaction levels.

Multicollinearity & Feature Reduction

- Multicollinearity observed across 'C,' 'D,' and 'V' feature groups.
- **'C1-C14'**: Counting features (e.g., associated addresses with payment cards).
- **'D1-D15'**: Timedelta features (e.g., days between previous transactions).
- **'Vxxx'**: Vesta's rich engineered features (ranking, counting, etc.).
- Approach: Applied PCA separately to each feature group.
- Goal: Reduce multicollinearity while retaining predictive power.
- Utilized Random Forest Feature Importance to focus on predictive features.
- Reduced features from 434 to improve model efficiency and interpretability.

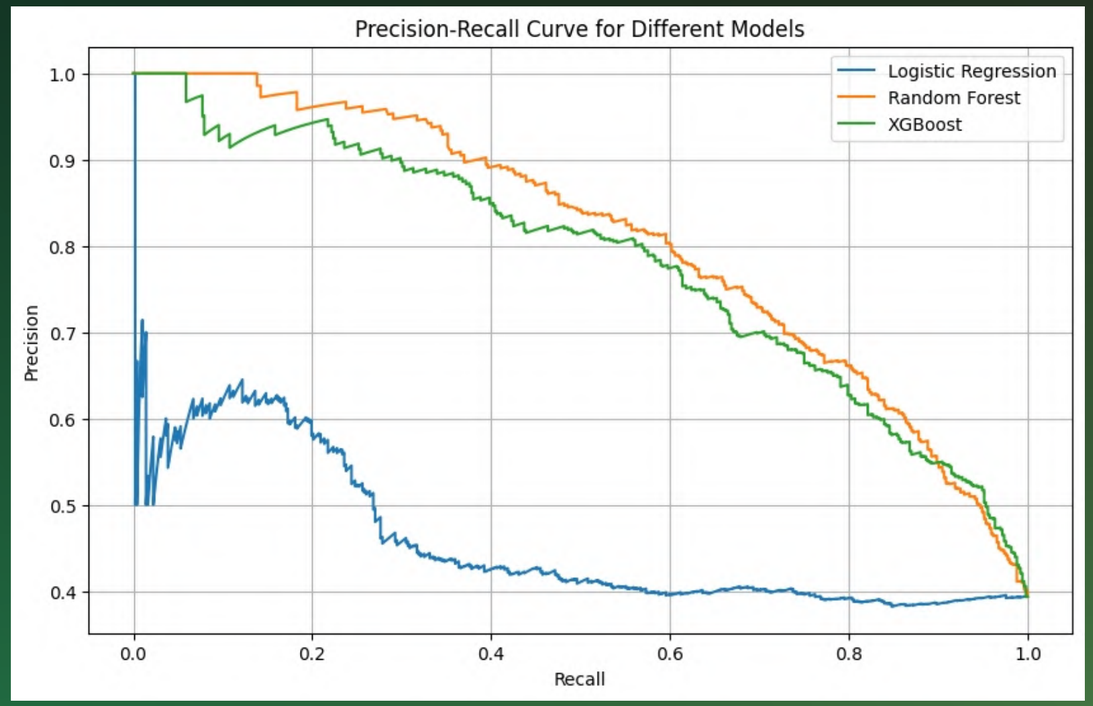


TOP 39 FEAUTRES FROM RANDOM FOREST FEATURE IMPORTANCE

PRECISION-RECALL CURVE ANALYSIS

Modeling

- Three models compared: Logistic Regression, Random Forest, and XGBoost.
- Logistic Regression exhibits consistently lower precision across recall levels.
- Random Forest and XGBoost exhibit similar, strong precision levels, especially at higher recall values.
- Both models demonstrate excellent ROC AUC scores, with Random Forest achieving 0.89 and XGBoost achieving 0.88.



XGBoost vs. Random Forest - Threshold Analysis

- XGBoost and Random Forest exhibit similar precision and recall at intersection points.
- XGBoost displays slightly higher precision at a lower threshold (0.3), indicating better high-recall performance.
- XGBoost shows a gradual recall decline with increasing thresholds, implying greater stability.
- **Our objective**: A model with stability and equilibrium between recall and precision across thresholds.
- XGBoost excels, especially at threshold 0.43, achieving a balanced recall-precision equilibrium.
- XGBoost emerges as a robust choice for our fraud detection objectives.

XGBoost

Threshold	*Precision (Fraud)*	*Recall (Fraud)*
0.3	*0.70*	*0.82*
*0.43****	*0.76*	*0.76*
0.5	*0.79*	*0.73*
0.7	*0.85*	*0.62*

Random Forest

Threshold	*Precision (Fraud)*	*Recall (Fraud)*
0.3	*0.66*	*0.88*
*0.45****	*0.75*	*0.76*
0.5	*0.79*	*0.71*
0.7	*0.91*	*0.48*

***INTERSECTION OF PRECISION AND RECALL

CONCLUSION

Final Words

Given the cryptic nature of the dataset, initial comprehension and feature engineering posed challenges. Every feature proved valuable, considering the issues of missing data and target variable imbalance. PCA and Random Forest Feature Importance helped immensely in reducing the large amount of features and place focus on those of greater importance.

The choice of threshold, depending on the bank's objectives, plays a pivotal role in balancing recall and precision.

However, in the future a significant enhancement lies in combining the strengths of both Random Forest and XGBoost. Such an ensemble approach would substantially boost the model's predictive capabilities, making it a compelling solution.