# Report on IEEE-CIS Fraud Detection Dataset

# Overview of the Dataset

The IEEE-CIS Fraud Detection dataset, provided by Vesta, contains 590,540 rows and 434 anonymized features, aiming to predict fraudulent transactions (isFraud). It is derived from real-world e-commerce transaction data, offering rich but anonymized features like device types, transaction amounts, and email domains. Given the imbalanced target variable (only 3.5% fraudulent transactions), the dataset requires advanced techniques for exploration, feature engineering, and modeling to ensure the effective detection of fraud while minimizing false positives and negatives.

* **Training Set:**
  + Contains 590,540 transactions.
  + Includes both labeled and unlabeled data.
* **Test Set:**
  + Contains 506,691 transactions.
  + The target (isFraud) is not provided.
* **Target Variable:**
  + isFraud: Binary classification label (1 = Fraud, 0 = Genuine).
* **Feature Categories:**
  + **Transaction Features:**
    - TransactionID, TransactionDT, TransactionAmt, and other transaction-specific information.
  + **Identity Features:**
    - Email domains, device information, and browser metadata.
  + **Categorical and Numerical Features:**
    - Variables are mixed, including encoded categorical features.
* **Time Dimension:**
  + TransactionDT is an incremental time measure but does not directly map to real-world time.

## Business Understanding

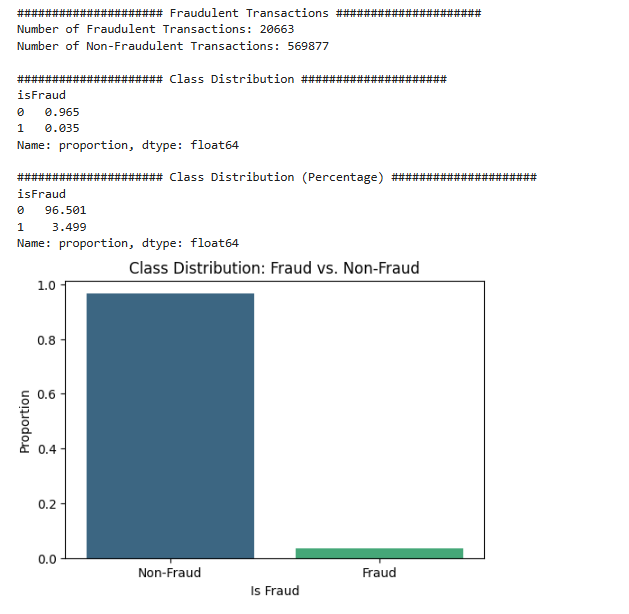
The primary objectives for this project are:

1. Fraud Detection: Accurately flagging fraudulent transactions while minimizing false positives to avoid blocking legitimate users.  
2. High-Risk Transactions: Applying stricter thresholds for high-value transactions to prevent significant losses.  
3. Quick and Scalable Models: Building computationally efficient models like LightGBM for real-time applications.  
4. Feature Insights: Identifying key fraud indicators to enhance future data collection and analysis.  
5. Balancing Security and Customer Convenience: Avoiding over-blocking, which could harm user experience.  
Metrics like AUC-ROC and F1-Score are emphasized to balance the competing priorities of false positives and false negatives effectively.

## EDA

**Target Distribution (isFraud)**

The target is highly imbalanced, with only **3.5%** of transactions being fraudulent. This imbalance can cause models to favor the majority class, leading to poor fraud detection.



**TransactionDT** feature represents time, likely in seconds, with values ranging from **86,400 to 15,811,131**. New features like **hour** and **day** were created for better model insights.

* **Fraud Trends by Hour**:
  + Higher fraud rates from **4 am to 12 pm**, peaking between **7 pm and 10 pm**.
  + Lowest fraud rates from **2 pm to 4 pm**.
  + This reflects a reversed trend: fraud increases when transaction frequency drops, likely due to people being less vigilant at these times.

A graph of green bars

Description automatically generated with medium confidence

**TransactionAmt** feature represents the transaction amount. A **log transformation** was applied to reduce the skew caused by very large transactions.

* The mean transaction amount for fraudulent transactions is higher than for non-fraudulent ones.
* **Extremely low** and **extremely high** transaction amounts are more likely to be fraudulent.

**Dist1 and Dist2**

These features might represent the distance between the transaction location and the card owner's home or work address.

**ProductCD**

A graph with blue and orange bars

Description automatically generated

the graph above, we observe that if a transaction is fraudulent, there is a 40% chance it belongs to Product C (though the reverse cannot be inferred). Given that only 10% of legitimate transactions are associated with Product C, the likelihood of fraud increases significantly when the product is C

**Addr1 and Addr2 - Masked Billing Region and Country**

The addr2 feature likely represents billing countries, with around 70 unique values. Country code **87** accounts for approximately **88%** of transactions, suggesting it likely represents U.S. accounts, as Vesta, the data provider, is primarily based in the United States.

**P\_emaildomain and R\_emaildomain - Purchaser and Recipient Email Domains**

P\_emaildomain represents the purchaser's email domain, while R\_emaildomain represents the recipient's. To simplify these features, similar domains were grouped (e.g., yahoo.com and yahoo.com.mx as "Yahoo"). Domains with fewer than 500 entries were labeled as "Others," and missing values were filled with "NoInf."

**Devicename-**Motorola and "Others" devices show the highest fraud rates, while "unknown\_device," though the most common, has a low fraud percentage. This suggests that fraudsters may prefer specific, less common devices.

A screenshot of a graph

Description automatically generated

## Key Insights from EDA

1. Fraudulent Transactions Exhibit Distinct Patterns:  
   - Transaction Time: Fraud is more likely during low transaction frequency periods (e.g., late night and early morning).  
   - Transaction Amount: Extremely low or high amounts are more likely to be fraudulent. Log transformation helped reduce skew and improve model performance.
2. Product and Device Features Matter:  
   - Fraud rates are significantly higher for Product C, suggesting fraudsters' preference for this product.  
   - Fraudsters tend to use less common devices like Motorola, as opposed to 'unknown\_device,' which is the most common but has low fraud rates.
3. Geographic and Email Domains Provide Clues:  
   - Grouping email domains improved model interpretability and accuracy.  
   - Address features (e.g., addr1, addr2) helped localize fraudulent transactions geographically.

## Summary of Feature Engineering (FE)

1. **Transaction Date Processing (TransactionDT)**:
   * Converted TransactionDT into date-related features (NewDate, NewDate\_YMD, NewDate\_YearMonth, etc.) to capture patterns based on time, such as day of the week and hour of the day.
2. **Transaction Amount Features (TransactionAmt)**:
   * Created New\_Cents to capture the fractional part of the transaction amount.
   * Binned TransactionAmt into 15 quantiles to group transactions by range (New\_TransactionAmt\_Bin).
3. **Card Feature Processing**:
   * Filled missing values in card-related features (e.g., card2, card3, card4, etc.) using the mode within each card1 group.
   * This aims to fill missing values based on common values within each card grouping.
4. **Email Domain Processing (P\_emaildomain, R\_emaildomain)**:
   * Grouped popular email domains into categories (Google, Yahoo, Microsoft) and labeled rare domains as "Others".
   * Filled missing values with "Unknown" to handle missing data effectively.
5. **Device Information Processing (DeviceInfo)**:
   * Standardized device information by mapping common device names (e.g., Samsung, Motorola, etc.) and grouping infrequent devices as "Others".
   * Missing values were filled with "unknown\_device".
6. **V-Features Processing (V1 - V339)**:
   * Grouped V features based on similar missing value patterns.
   * Applied **PCA** (Principal Component Analysis) and **MinMax Scaling** within each group to create compact features, reducing dimensionality.
7. **Combined Features**:
   * Created combined features such as New\_card1\_card2, New\_addr1\_addr2, and New\_card1\_card2\_addr1\_addr2 to capture interactions between different cards and addresses.
8. **Aggregation Features**:
   * Created features based on the mean and standard deviation of TransactionAmt, id\_02, and D15 within groups defined by card1, card4, and addr1.
   * These features help capture outliers by comparing transaction values to group averages.
9. **Frequency Encoding**:
   * Applied frequency encoding to selected categorical features (New\_TransactionAmt\_Bin, card4, card6, P\_emaildomain, R\_emaildomain, DeviceType, DeviceInfo) to reflect the frequency of each category.
10. **Dropping Unnecessary Features**:

* Dropped columns that were transformed or replaced by derived features to keep the dataset compact.

1. **Label Encoding**:

* Applied label encoding to remaining categorical features, converting them to numerical codes for compatibility with machine learning models.

## Machine Learning Models:

- LightGBM, XGBoost, CatBoost: These gradient-boosted decision tree algorithms are chosen for their ability to handle large, imbalanced datasets with minimal preprocessing and excellent scalability.  
- Random Forest: Provides robust feature importance insights and performs well in a variety of tasks, though it may require more computational resources.

The final LightGBM model showcased strong performance with a **Training Accuracy of 99.20%** and a **Validation Accuracy of 98.40%**, indicating excellent generalization. It achieved a **Precision of 94.89%**, ensuring most flagged transactions were truly fraudulent, and an **AUC of 0.9655**. Submitting the model to Kaggle gave a score of **0.942167**, which I am happy with as a student. This project was a great learning experience, and I plan to improve the model further to detect more fraud cases while keeping it accurate.

A blue and white graph

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1-Score** | **AUC** |
| LGBMClassifier | 0.948863 | 0.572198 | 0.713871 | 0.965531 |
| XGBoost | 0.95156 | 0.581913 | 0.722172 | 0.96445 |
| RandomForest | 0.958896 | 0.482154 | 0.641656 | 0.940508 |
| CatBoost | 0.940285 | 0.471754 | 0.628277 | 0.940107 |