**Credit Card Transactions – Fraud Detection**

Akshay Sharma

Department of Data Science, Bellevue University

DSC530-T303: Data Exploration & Analysis

Matthew Metzger

March 01, 2025

## **Introduction**

Credit card fraud is a growing concern for financial institutions, businesses, and consumers. The increasing adoption of digital transactions has provided fraudsters with sophisticated ways to exploit vulnerabilities in payment systems. Traditional rule-based fraud detection techniques are often inadequate in handling evolving fraudulent behaviors. This study explores credit card fraud detection using machine learning techniques, focusing on dataset analysis, feature engineering, model development, and evaluation.



## **The Problem Statement**

The primary statistical question in this analysis was: **Do transaction amount, city population, transaction time, or merchant category significantly impact the likelihood of fraud in credit card transactions?** This question guided our exploration into transaction behaviors and patterns associated with fraudulent activities. We aimed to identify significant predictors of fraud through hypothesis testing and exploration data analysis (EDA).

## **Data For Analysis**

Dataset which has been identified for this study:

1. **Credit Card Transactions Fraud Detection Dataset**:
   * **Source**: Kaggle - <https://www.kaggle.com/datasets/kartik2112/fraud-detection>
   * **Description**: A simulated dataset containing legitimate and fraudulent transactions from January 1, 2019, to December 31, 2020. It includes features such as transaction time, amount, location, and merchant details.

## **Data Preparation and Cleaning**

To analyze and process this dataset, the following steps were performed:

* **Data Cleaning:** No missing values were found, but potential duplicate transactions were examined.
* **Exploratory Data Analysis (EDA):**
  + Class distribution was visualized to highlight the imbalance in fraud cases.
  + Transaction amount distributions showed that fraudulent transactions tend to have unique spending patterns.
  + Time-based analysis revealed that fraud often occurs at unusual hours.
  + Correlation analysis helped in identifying relationships between features.

## **Outcome of Exploratory Data Analysis (EDA)**

The EDA revealed several key findings:

* **Transaction Amount:** Fraudulent transactions tend to have higher average amounts compared to non-fraudulent ones. A ‘t-test’ confirmed this difference was statistically significant.
* **City Population:** Fraud was more common in highly populated cities, indicating a potential risk concentration in urban areas.
* **Transaction Time:** Fraudulent transactions frequently occur during unusual hours, possibly to avoid detection.
* **Merchant Category:** Certain merchant categories had a disproportionate occurrence of fraud, as confirmed by a Chi-square test.
* **Demographics:** Fraud cases were more frequent among younger individuals (ages 25-40) and male cardholders. Additionally, fraud was less common in high-income groups but prevalent among middle-income individuals.
* **Distribution Analysis:** Probability Mass Function (PMF) and Cumulative Distribution Function (CDF) analyses demonstrated a heavy-tailed distribution of fraudulent transactions, with fraudsters conducting fewer but high-value transactions.
* **Machine Learning Insights:** A Logistic Regression model achieved a 95.34% accuracy rate, improving fraud detection. A confusion matrix analysis showed a significant reduction in false negatives.

## **Missed Aspects in Analysis**

Despite the thorough investigation, several aspects could have been explored further:

* **Device Type:** Identifying fraud trends across different devices (mobile, desktop, etc.) could provide valuable insights.
* **Customer Behavior:** Analyzing repeat fraud attempts by the same individuals could enhance fraud prevention measures.
* **Merchant Risk Scores:** Some merchants may be historically linked to fraudulent activities, and incorporating risk scores could refine fraud detection models.
* **Geolocation Data:** Longitude and latitude anomalies could reveal fraudulent patterns, and cross-verifying customer-merchant distances may help detect suspicious activity.

## **Additional Variables That Could Have Helped**

Incorporating **IP address tracking, transaction velocity (number of transactions per unit time), and user login behavior** could have improved the model’s ability to detect fraud. Additionally, **geographical distance analysis** between the merchant and customer locations could help flag improbable transactions.

## **Incorrect Assumptions & Challenges Faced**

Several challenges and potential incorrect assumptions were encountered:

* **Class Imbalance:** Fraud cases were significantly outnumbered by non-fraud cases, leading to model bias. We applied to **SMOTE (Synthetic Minority Over-sampling Technique)** to mitigate this issue.
* **Normality Assumption:** Some hypothesis tests assumed a normal distribution of transaction amounts, which was not the case. Alternative non-parametric tests like the **Mann-Whitney U test** were used to address this.
* **Changing Fraud Patterns:** Fraud behavior evolves over time. The model was trained on static historical data, which may not fully generalize about future fraud patterns.
* **Feature Engineering Complexity:** Identifying non-linear relationships and interactions between variables required deeper analysis beyond traditional statistical methods.

## **Visualization Insights**

* **Transaction Amount Distribution:** Fraudulent transactions tend to occur at specific price points and have higher values than non-fraudulent transactions.
* **Fraud by Merchant Category:** Some merchant categories experience disproportionately higher fraud occurrence, as seen in the PMF analysis.
* **City Population & Fraud:** Higher fraud rates were observed in urban locations, with cities over 500,000 residents experiencing the most fraud.
* **Scatter Plot Analysis:** Fraud clusters were identified when comparing **transaction amounts, city population, and geographic locations**.

## **Conclusion**

This study successfully identified key fraud predictors using EDA, hypothesis testing, and machine learning models. The results confirmed that **transaction amount and merchant category significantly impact fraud occurrence**. The **95.34% accurate fraud detection model** demonstrated improvements but still faced challenges in reducing false negatives. Future work should explore **advanced anomaly detection techniques, additional behavioral features, and improved real-time fraud detection models** to further enhance fraud prevention strategies.