Fraud Detection in Credit Card Transactions

Machine Learning Approach to Detect Fraudulent Transactions

Shivangini Marjiwe

March 2025

Introduction

- Credit card fraud is a major concern for financial institutions.
- Fraudulent transactions cause significant financial losses.
- **Goal:** Develop a machine learning model to detect fraud effectively.

- **Objective:** Identify fraudulent transactions while minimizing false positives.
- **Challenges:**

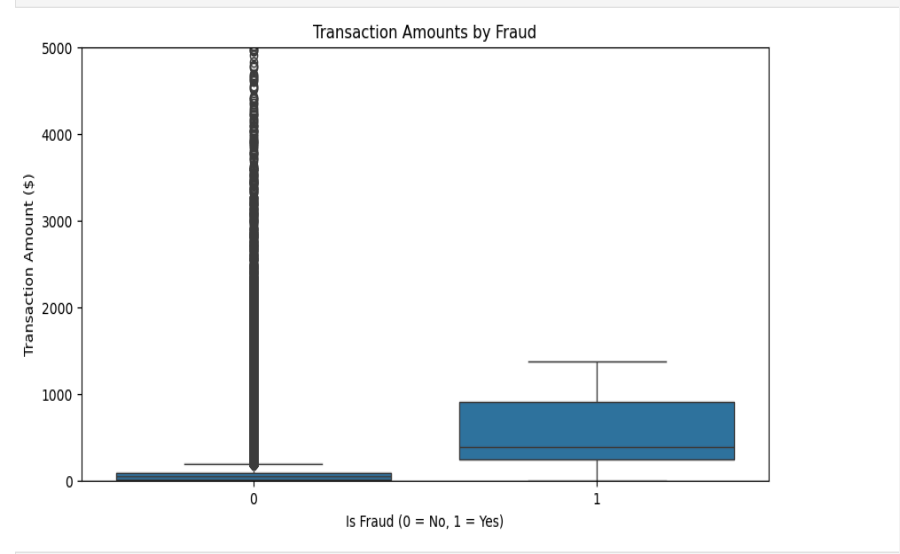
    1. Fraud transactions are rare (imbalanced dataset).
    2. Need high precision & recall for fraud detection.
    3. Performance Metrics: Precision, Recall, F1-Score, ROC AUC.

Data Cleaning

- **Removed:** Missing values, duplicates, and unnecessary features.
- **Dropped High-Cardinality Features:** Street, city, zip, job, trans_num.
- **Converted Data Types:** Optimized memory usage.

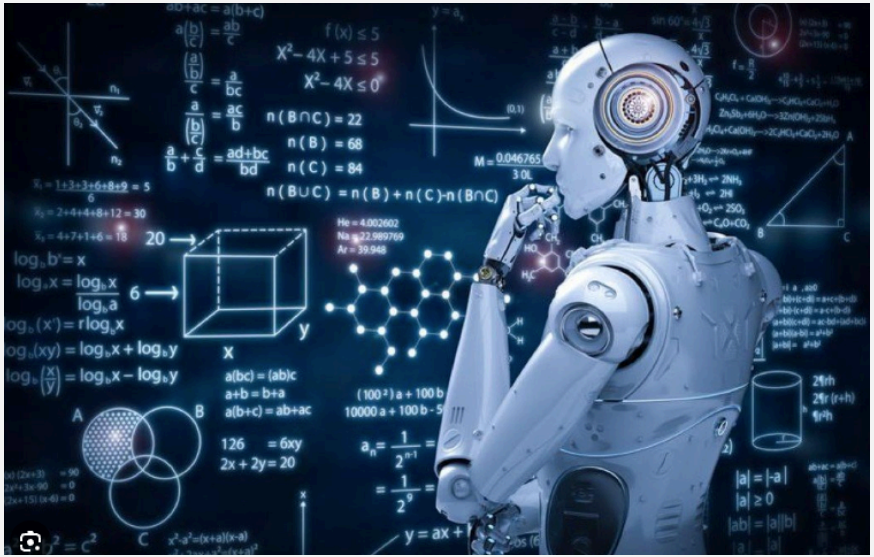
Exploratory Data Analysis (EDA)

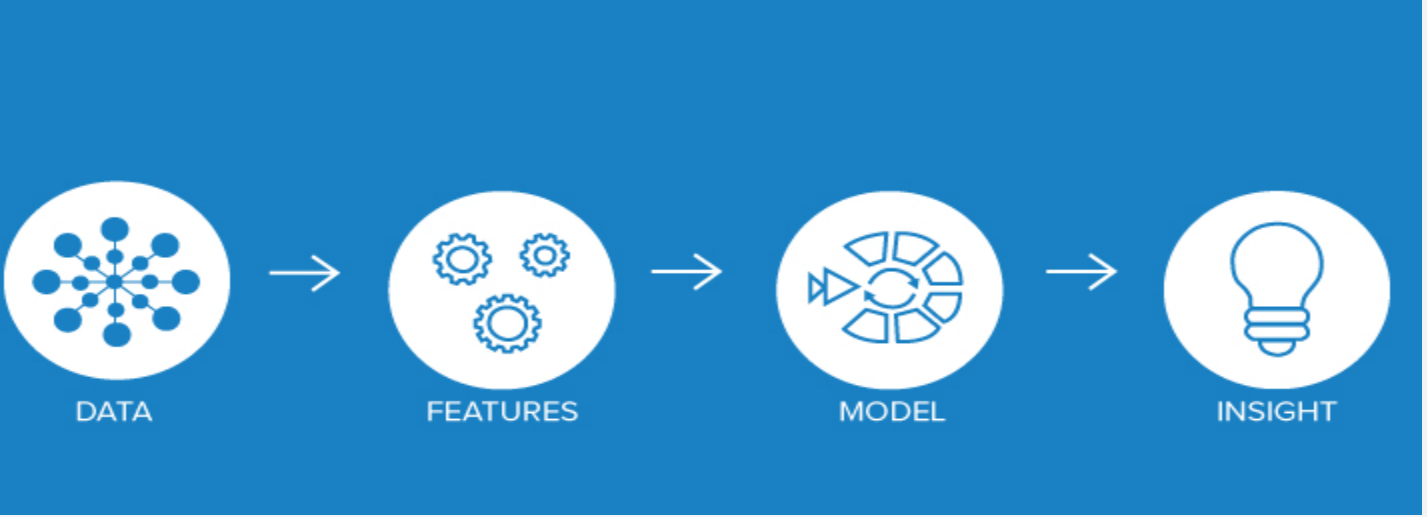
- Fraud vs. Non-Fraud Transactions Distribution.
- **Transaction Amounts:** Fraud transactions tend to have higher amounts.
- **Transaction Time Patterns:** Fraudulent transactions peak at certain hours.
- **Merchant Category Analysis:** Some categories have more fraud cases



Preprocessing

- **One-Hot Encoding:** Applied to categorical variables.
- **Feature Scaling:** Standardized numerical features.
- **Train-Test Split:** Stratified split to maintain class balance.
- **Sparse Format:** Categorical features converted to save memory.

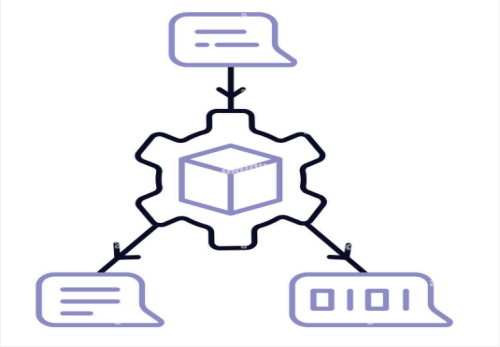




Feature Engineering

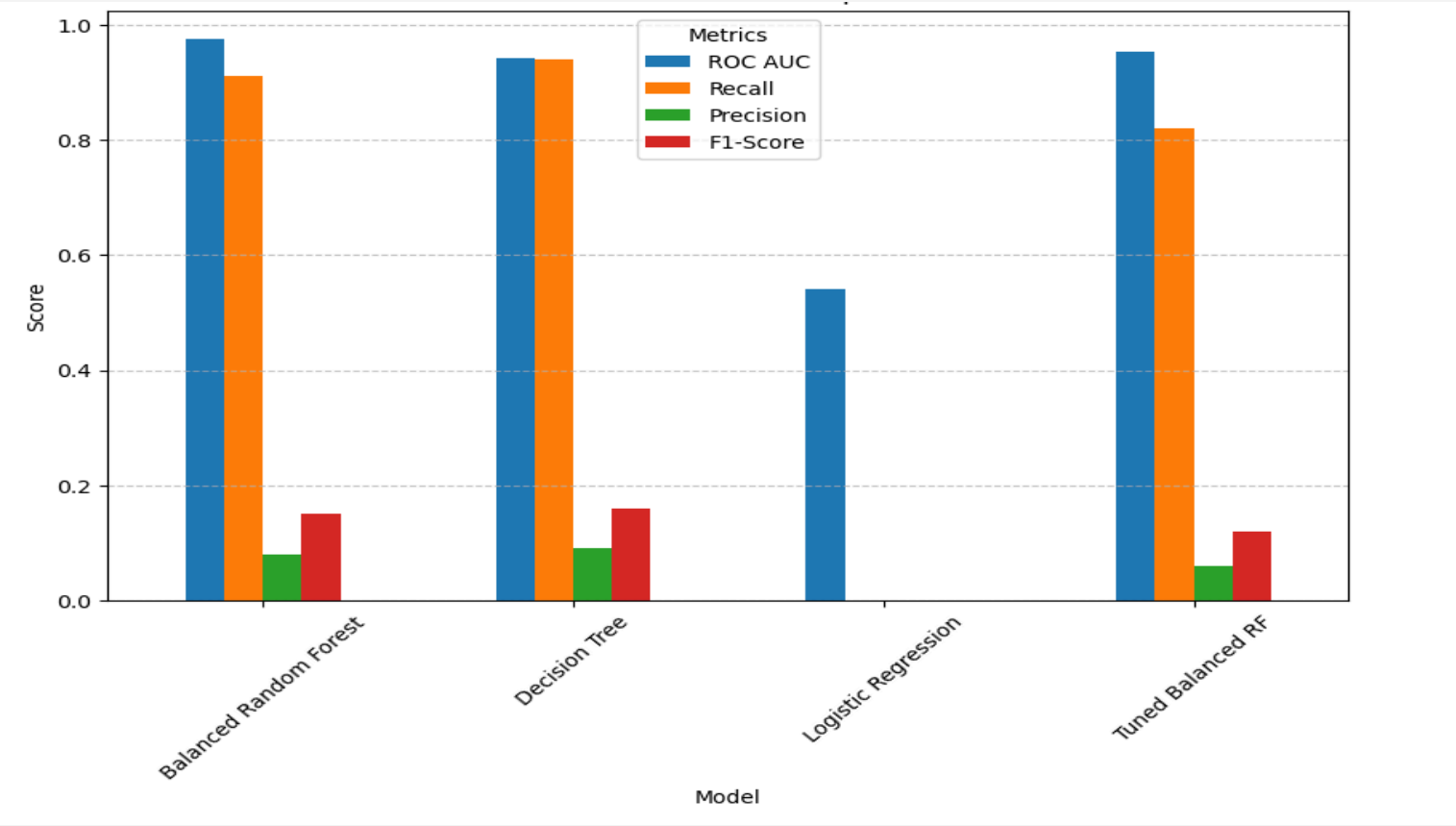
- **Extracted Time-Based Features:** Transaction hour, day of the week.
- **Encoded Merchant Categories:** One-hot encoding.
- **Normalized Continuous Features:** Transaction amount, latitude, longitude.
- **Final Features:** Amt, city_pop, age, trans_hour, lat, long, merch_lat, merch_long.

Trained Models



Model	ROC AUC	Recall	Precision	F1-Score
Balanced Random Forest	0.9750	0.91	0.08	0.15
Decision Tree	0.9421	0.94	0.09	0.16
Logistic Regression	0.5403	0.00	0.00	0.00
Tuned Balanced RF	0.9521	0.82	0.06	0.12

Model Performance Comparison



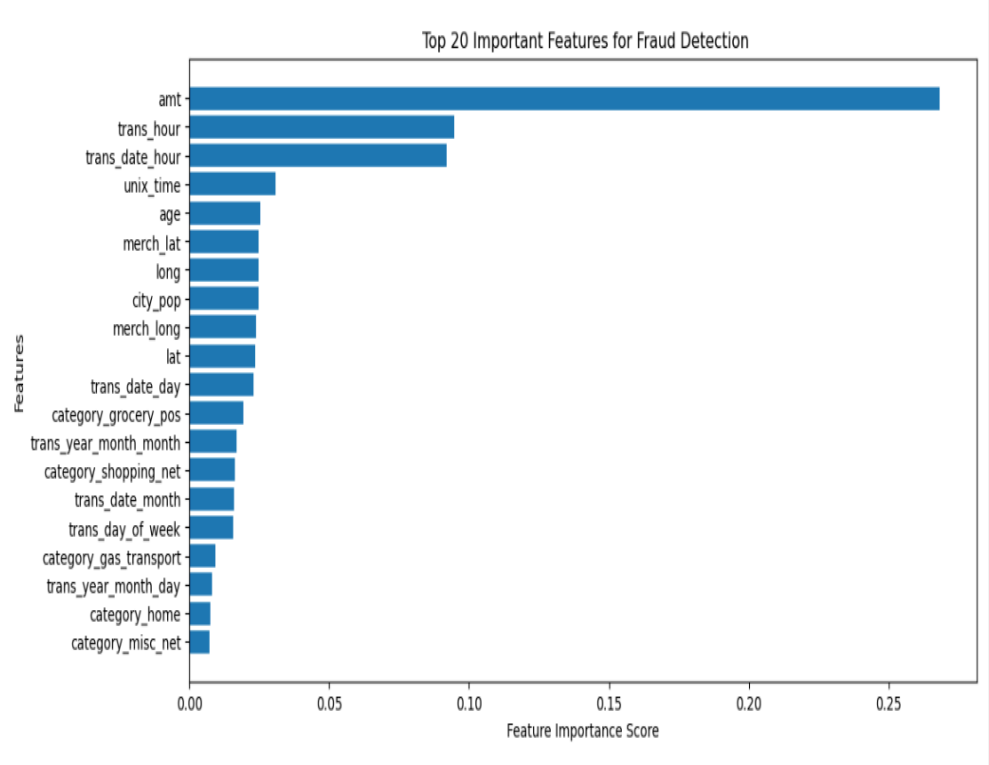
Model Selection

- **Best Model:** Balanced Random Forest (BRF)
- **Why?**
  - Highest ROC AUC (0.9750) and strong recall (0.91).
  - Best balance between fraud detection and false positives.
- **Why Not Tuned BRF?**
  - Slightly worse recall than default BRF.
- **Why Not Logistic Regression?**
  - Completely ineffective with 0 recall.

1. Balanced Random Forest
2. Decision Tree
3. Logistic Regression
4. Tuned Balanced RF

Feature Importance Analysis

- **Top 5 Important Features for Fraud Detection (Balanced RF):**
  1. Transaction Amount (amt)
  2. Transaction Hour (trans_hour)
  3. Transaction Date Hour (trans_date_hour)
  4. Unix Time (unix_time)
  5. Age (age)
- Location-based features (latitude/ longitude) had lower importance.



Conclusion & Recommendations

Findings:

- Balanced Random Forest is the best model.
- Fraud occurs more in high-amount transactions and specific time periods.

Recommendations:

1. Implement BRF Model in real-time fraud detection systems.
2. Enhance Fraud Prevention Measures during peak fraud hours.
3. Monitor High-Value Transactions closely.

Thank You!

Any questions?