Credit Card Fraud Detection Model

Executive Summary

Problem Statement

Credit card fraud poses significant challenges, with global fraud losses exceeding $32 billion in 2023. Fraudulent transactions constitute only a very small percentage of all credit card activities, making accurate detection difficult. Businesses face dual challenges: identifying fraud effectively and minimizing false positives, which impact customer experience and operational costs.

Current Solutions

Traditional fraud detection methods struggle to balance fraud identification with operational efficiency. False positives increase manual reviews and frustrate customers, while undetected frauds cause significant financial loss. Existing systems often lack scalability and real-time capabilities for large datasets.

Our Approach

1. **Data and Insights**:

   - Dataset: 1.29M transactions spanning 1.5 years, with detailed features including transaction details (amount, time), user demographics, and merchant information.
   - Exploratory analysis highlighted extreme data imbalance (0.58% frauds), requiring specialized strategies.

2. **Pipeline Implementation**:

   i. **Data Ingestion & Exploration**
   ii. **Feature Engineering**: Developed advanced features like:
      - Age
      - Transaction Hour
      - Transaction Month
      - Transaction Year
      - Transaction Day of Year
      - Day of Week
      - Distance between User and Merchant
      - Merchant Popularity
      - Mean Transaction Amount per User
      - Transaction Amount Deviation from Mean
      - Transaction Count per User
      - Fraud Rate by Location
      - Transaction Patterns like if it is Recurring

  - Age Group
  - Time of Day
  - Mean of Distance between User & Merchant

iii. **Dropped PII and Irrelevant Columns**

iv. **Splitting Data in Train, Validation & Test**
  - Stratified Split between Train & Validation to balance the imbalanced data

v. **Imputation:**
  - Imputation wasn't necessary as we dropped unnecessary column with null values

vi. **Encoding:**
  - Multiple Encoding methods were tested

vii. **Scaling:**
  - Standard Scalar was used to scale data for regression models

viii. **Feature Tuning – Multiple feature tuning methods were used**
  - XGBoost – Feature Importance
  - AdaBoost – Feature Importance
  - Correlation Matrix – to review Multi-Collinearity/Redundancy

ix. **Metrics for Evaluation**:
  - **Primary Metric**: Recall, emphasizing fraud detection accuracy.
  - **Supplementary Metric (Custom)**:
    - Weighted Cost-Aware Accuracy (WCAA), quantifying the financial impact of undetected fraud and false positives.
    - Cost of Model

x. **Model Selection and Training**:
  - **Multiple Models were trained:**
    - LogisticRegression
    - RandomForest
    - XGBClassifier
    - CatBoost
    - LightGBM
    - AdaBoost
    - IsolationForest
    - Stacked – XGBClassifier, BalancedRandomForest, LogisticRegression
    - Stacked – RandomForestClassifier, LinearSVC, CatBoost, AdaBoost, HistGradientBoostingClassifier, LogisticRegression
    - Keras
  - **Feature Tuning – Multiple feature tuning methods were used**
    - XGBoost – Feature Importance
    - AdaBoost – Feature Importance
    - Correlation Matrix – to review Multi-Collinearity/Redundancy
  - **Hyperparameter Tuning**
    - GridSearch
    - RandomSearch
  - **Advanced Techniques for imbalanced Dataset Optimization**
    - SMOTE – Synthetic Minority Oversampling Technique
    - Class Weighting
    - Focal Loss

3. **Results**: shown in presentation
    a. IsolationForest Algorthim showed the best results for our highly imbalanced dataset.
    b. Many other algorithms showed great potential and would require further time to optimize

Business Implications

- **Fraud Detection**: The model's high recall significantly reduces undetected fraud.
- **Customer Experience**: Controlled false positives maintain customer satisfaction.
- **Cost Reduction**: Quantified savings using custom metrics - WCAA ($ Value), and Cost of Model demonstrate clear business benefits.

Next Steps

1. **Real-Time Integration**:
    - Deploy APIs for live transaction monitoring.
    - Ensure scalable architecture for high-velocity data.
2. **Enhanced Visualizations**:
    - Develop customizable dashboards for fraud trends.
3. **Generalization**:
    - Adapt the pipeline for other datasets and industries.
4. **Continuous Improvement**:
    - Use adaptive learning for evolving fraud patterns.
    - Explore alternative metrics like Weighted Gini or Amex Metric for nuanced insights.

Conclusion

This project underscores the potential of machine learning in addressing credit card fraud challenges. By achieving high recall and quantifying financial impacts with Custom Metrics, we deliver a solution that reduces fraud while preserving customer trust. Our roadmap for real-time implementation and scalability ensures alignment with business needs and future growth opportunities.