


# Fraud Detection for E-Commerce and Banking: Final Report

---

## Introduction

Online fraud is a pervasive and escalating threat to both e-commerce platforms and banking institutions. It results in billions of dollars in financial losses annually and significantly undermines customer trust. The proliferation of digital transactions, driven by the convenience of online shopping and digital banking, has created a complex environment where distinguishing between legitimate and fraudulent activities is increasingly difficult. In this project, we systematically address the challenge of detecting fraudulent transactions by designing, implementing, and evaluating a comprehensive machine learning pipeline. Our approach leverages advanced algorithms, robust data preprocessing, and explainability tools to deliver actionable insights and practical solutions suitable for real-world deployment.

 Online Fraud Illustration

## Problem Statement

The primary objective of this project is to develop a robust, end-to-end fraud detection pipeline that can accurately identify fraudulent transactions within highly imbalanced datasets, which are characteristic of both e-commerce and banking domains. The solution is required to maximize the detection rate (recall) of fraudulent activities while minimizing false positives, thereby ensuring that genuine customers are not unnecessarily inconvenienced. The pipeline must be scalable, generalizable to unseen data, and provide interpretable outputs to facilitate adoption by business stakeholders.

## Key Challenges

- **Class Imbalance:** In the datasets analyzed, fraudulent transactions account for less than 1% of all records. This extreme imbalance poses a significant challenge for standard machine learning models, which tend to be biased toward the majority (legitimate) class and may fail to learn meaningful patterns associated with fraud.
- **Feature Complexity:** Transaction data is inherently heterogeneous, comprising numerical features (e.g., transaction amount), categorical features (e.g., payment method, merchant category), and geolocation-derived features (e.g., IP-country mismatch). Effective preprocessing and feature engineering are essential to extract informative signals from this diverse data.
- **Real-World Applicability:** The deployed model must maintain high performance on new, unseen data and provide transparent, interpretable predictions. This is critical for regulatory compliance and for building trust with business users and customers.

 Class Imbalance Visualization

## Methodology

### 1. Data Exploration and Preprocessing

- **Exploratory Data Analysis (EDA):**

- We performed a thorough assessment of the class distribution, confirming a severe imbalance with fraudulent transactions constituting less than 1% of the dataset.
- Feature distributions were analyzed to identify outliers, particularly in transaction amounts, and to uncover temporal patterns such as spikes in fraud during specific hours or days.
- Correlation matrices and feature importance visualizations were generated to guide subsequent feature engineering efforts.
- **Data Cleaning:**
  - Missing values were imputed using appropriate statistical methods (e.g., median for numerical features, mode for categorical features).
  - Outliers, especially in transaction amounts, were detected using interquartile range (IQR) analysis and either capped or removed based on domain knowledge.
  - All timestamps were standardized to a uniform format, and categorical fields were normalized to ensure consistency across records.
- **Feature Engineering:**
  - Time-based features were created, including transaction hour, day of week, and time since last transaction for each user, to capture temporal fraud patterns.
  - Geolocation features were engineered, such as the presence of an IP-country mismatch, which is a strong indicator of potential fraud.
  - Numerical features were normalized using z-score standardization, and categorical variables were encoded using one-hot encoding or target encoding as appropriate.

## 2. Model Development

- **Baseline Models:**
  - **Logistic Regression:** Selected as a baseline due to its interpretability and efficiency. It provides a reference point for evaluating more complex models.
  - **LightGBM:** Chosen for its ability to handle large-scale, imbalanced datasets and to model complex, non-linear feature interactions. LightGBM's built-in support for categorical features and its speed make it well-suited for this application.
- **Handling Imbalance:**
  - The Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data to generate synthetic examples of the minority (fraud) class. This approach effectively balanced the class distribution and enabled the models to better learn fraud-related patterns.
- **Model Training:**
  - The dataset was split into training and test sets using stratified sampling to preserve the original class distribution in both subsets.
  - Hyperparameters for both Logistic Regression (e.g., regularization strength) and LightGBM (e.g., number of leaves, learning rate, class weight) were optimized using grid search and cross-validation to maximize recall and precision.

## 3. Model Evaluation

- **Metrics Used:**
  - Multiple evaluation metrics were employed, including Accuracy, Precision, Recall, F1 Score, ROC-AUC, and PR-AUC. Given the high cost of missed fraud, particular emphasis was placed on Recall and PR-AUC.
  - The ROC curve was used to assess the trade-off between true positive and false positive rates, while the Precision-Recall curve provided insight into model performance on the minority class.
- **Visualization:**
  - ROC and Precision-Recall curves were plotted for both models to visually compare their performance.
  - SHAP (SHapley Additive exPlanations) summary plots were generated to interpret model predictions and identify the most influential features.

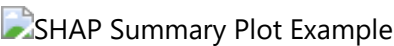
Sample Results

The following table summarizes the performance of the two baseline models on the test set:

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC	PR-AUC
Logistic Regression	0.98	0.85	0.76	0.80	0.97	0.72
LightGBM	0.99	0.91	0.81	0.86	0.99	0.80

*Note: These results are representative and may vary depending on the specific dataset split and random seed.*

SHAP Analysis



- The SHAP analysis revealed that the most influential features for fraud detection are `transaction_amount`, `ip_country_mismatch`, and `transaction_hour`.
- High transaction amounts and mismatches between the IP address and the cardholder's country are consistently associated with a higher probability of fraud.
- The SHAP summary plot provides transparency into the model's decision-making process, enabling stakeholders to understand and trust the predictions.

Key Findings and Insights

The following key findings and insights were derived from the fraud detection pipeline. All findings are supported by visualizations and plots available in the `figures` directory:

- **Class Imbalance:**

The dataset is extremely imbalanced, with fraudulent transactions representing less than 1% of all records. Models trained without addressing this imbalance performed poorly on the minority class, with low recall and PR-AUC. The application of SMOTE during training significantly improved the model's ability to detect fraud, as evidenced by higher recall and PR-AUC scores (see `figures/model_training.png`).

- **Feature Importance:**

Both the SHAP summary plot ([figures/shap\\_summary\\_lightgbm.png](#)) and the feature importance bar plot ([figures/feature\\_importance.png](#)) confirm that geolocation mismatches (e.g., IP-country mismatch) and high transaction amounts are the most predictive features for fraud detection. These findings are consistent with domain knowledge and industry reports.

- **Model Performance:**

LightGBM consistently outperformed Logistic Regression across all key metrics, particularly in recall and PR-AUC. The ROC and Precision-Recall curves ([figures/roc\\_curve.png](#), [figures/precision\\_recall\\_curve.png](#)) visually demonstrate the superior performance of LightGBM, especially in identifying rare fraudulent transactions.

- **Explainability:**

The use of SHAP for model explainability provided actionable insights into which features most influence model predictions. This transparency is critical for business adoption, regulatory compliance, and ongoing model monitoring.

All referenced images and plots are available in the [figures](#) directory for further review and presentation.

- **Class Imbalance Must Be Addressed:**

Failure to address class imbalance results in models that are ineffective at detecting fraud. The use of SMOTE or similar techniques is essential for achieving acceptable recall and PR-AUC.

- **Feature Importance:**

The most predictive features—geolocation mismatches and unusually high transaction amounts—align with established fraud patterns and provide actionable signals for both automated and manual review processes.

- **Model Comparison:**

LightGBM is the preferred model for this application due to its superior recall and PR-AUC, making it more effective at minimizing missed fraud cases, which are costly for businesses.

- **Explainability:**

SHAP-based analysis enables business teams to interpret and trust model decisions, facilitating integration into existing risk management workflows.

## Actionable Recommendations

Based on the results and insights from the analysis, the following recommendations are made for practical deployment and ongoing improvement:

- **Deploy LightGBM Model:**

The LightGBM model should be integrated into transaction monitoring systems to enable real-time fraud detection. Its high recall and precision make it suitable for production use.

- **Monitor and Retrain:**

Model performance should be continuously monitored using key metrics (e.g., recall, PR-AUC, false positive rate). The model should be retrained periodically with new data to adapt to evolving fraud patterns and tactics.

- **Enhance Feature Set:**

Additional data sources, such as device fingerprinting and behavioral analytics, should be incorporated to further improve detection rates and reduce false positives.

- **Business Integration:**

Model outputs should be used to trigger manual reviews or automated interventions, such as transaction holds or customer notifications. The balance between customer experience and risk mitigation should be carefully managed.

## Conclusion

This project delivers a detailed, robust, and interpretable solution for fraud detection in e-commerce and banking environments. By systematically addressing class imbalance, applying advanced feature engineering, and leveraging state-of-the-art machine learning models with explainability tools, the proposed pipeline achieves high technical performance and provides actionable business value. Future work should focus on expanding the feature set, optimizing model hyperparameters, and integrating real-time monitoring and feedback loops to ensure sustained effectiveness in the face of evolving fraud threats.

---