Fraud Detection Model using LightGBM - Project Documentation

1. Introduction

This project is focused on detecting fraudulent transactions using machine learning, specifically the LightGBM algorithm.

The dataset includes customer demographics and transaction data. The model handles imbalanced data using SMOTE and

evaluates performance using metrics such as Precision, Recall, F1 Score, and AUC-ROC.

2. Data Preprocessing

- Loaded and cleaned the dataset (removed quotes and dropped unused columns).
- Encoded categorical variables using Label Encoding.
- Split data into train and test sets with stratification to preserve class ratio.
- Balanced the training data using SMOTE to handle class imbalance.

3. Model Building

Used LightGBMClassifier with parameters optimized for binary classification:

- class_weight='balanced' to address imbalance
- learning_rate=0.02 and 2000 estimators
- max_depth=10, num_leaves=64

Model was trained with early stopping and evaluated on a validation set.

4. Threshold Optimization

Predicted probabilities were used to find the optimal threshold that maximizes the F1 score, which balances precision and recall.

This is important in fraud detection where false positives and false negatives have high impact.

5. Evaluation Metrics

After applying the optimal threshold, the following metrics were calculated:

- Precision: Measures how many predicted frauds were actually frauds.
- Recall: Measures how many actual frauds were caught.
- F1 Score: Harmonic mean of precision and recall.
- AUC-ROC: Area under the ROC curve, summarizing model discrimination.

6. Visualizations

Several visualizations were created to interpret the model performance:

- Feature Importance: Shows most influential features in decision making.
- Confusion Matrix: Summarizes prediction outcomes.
- Probability Distribution: Shows confidence of fraud predictions.
- Precision-Recall Curve: Visualizes tradeoff between precision and recall.
- ROC Curve: Visualizes tradeoff between sensitivity and specificity.

7. Conclusion

The LightGBM-based fraud detection model demonstrates strong performance on imbalanced data when combined with SMOTE and
threshold optimization. Visualizations provide valuable insights into the model's decision-making process.