IEEE-CIS Fraud Detection using Machine Learning

Abstract

This project focuses on detecting fraudulent online transactions using the IEEE-CIS Fraud Detection dataset. A complete machine learning pipeline was developed, including data preprocessing, feature engineering, model training, evaluation, and threshold optimization. LightGBM was used as the primary model due to its efficiency and strong performance on large, tabular datasets.

Dataset Description

The IEEE-CIS Fraud Detection dataset contains transaction and identity information. It is highly imbalanced, with fraudulent transactions accounting for approximately 3.5% of the data. The dataset includes anonymized numerical features, categorical variables, time-based features, and card-related attributes.

Methodology

The raw transaction and identity datasets were merged using TransactionID. Extremely sparse features were removed to reduce noise and memory usage. Categorical variables were encoded numerically, and a time-based train-validation split was used to prevent data leakage. LightGBM was selected as the classification model due to its ability to handle high-dimensional data efficiently.

Model Training

A baseline LightGBM model was trained using ROC-AUC as the evaluation metric. Feature importance analysis revealed that anonymized engineered features, count-based features, and time-related variables contributed most significantly to fraud detection.

Class Imbalance Handling

To address class imbalance, an alternative model using class-weighted loss (scale_pos_weight) was trained. Although this improved sensitivity to fraudulent transactions, the baseline model achieved higher ROC-AUC and was selected as the final model.

Evaluation Metrics

Metric	Value
Accuracy	0.9727
Precision	0.6443
Recall	0.4636

Threshold Tuning

Threshold tuning was performed to prioritize recall, as missing fraudulent transactions is more costly than false positives. A lower classification threshold improved recall while maintaining reasonable precision, demonstrating the importance of decision-level optimization.

Conclusion

This project demonstrates a complete and well-structured machine learning approach to fraud detection. By combining careful preprocessing, robust modeling, and thoughtful evaluation, the system effectively identifies fraudulent transactions while addressing real-world challenges such as class imbalance.