Fastag Fraud Detection

BY SANHITA SAXENA

Email: sanhitasaxena@gmail.com

This project focuses on leveraging machine learning classification techniques to develop an effective fraud detection system specifically for FASTag transactions. The dataset includes crucial features such as transaction details, vehicle information, geographical location, and transaction amounts. The primary objective is to build a robust model capable of accurately identifying instances of fraudulent activity, thereby safeguarding the integrity and security of FASTag transactions.

PROJECT OBJECTIVES

- Exploratory Data Analysis
- Data Preprocessing
- Model Selection and Comparison
- Model Training and Evaluation (Hyperparameter Tuning)
- Pipeline Building

EXPLORATORY DATA ANALYSIS

- Peak fraudulent activity times: Most frauds occur at 4 PM, 10 PM, and 6 AM.
- Months with highest fraud incidents: January recorded the highest number of frauds, followed by March.
- Lane with the highest fraudulent activity: Lane B102 experiences the most fraudulent transactions.
- Vehicle types involved in fraud: Large vehicles, particularly SUVs and Vans, are most frequently involved in fraudulent activities. Sedans and Trucks also show significant involvement.

DATA PREPROCESSING

- Dropped Columns: Removed 'Transaction_ID', 'FastagID', 'Vehicle_Plate_Number', 'Transaction_Amount', 'Amount_paid', and 'Timestamp' as they were not relevant for analysis.
- Feature Engineering: Created a 'State' column from 'Vehicle_Plate_Number' to map states accurately. Extracted 'Month' and 'Time of Day' from 'Timestamp'.
- Encoding: Used One-Hot Encoding (OHE) for categorical columns to handle categorical data effectively.
- Scaling: Applied StandardScaler to numeric columns for uniform scaling and preparation for machine learning models.

MODEL SELECTION AND COMPARISON

- **Models Evaluated:** Logistic Regression, Random Forest Classifier, KNN Classifier, Gradient Boosting Classifier, XGBoost, CatBoost, SVM Classification.
- **Performance Evaluation:** Compared using F1-score due to dataset imbalance, which balances precision and recall effectively.
- **Best Performing Model:** KNN Classifier achieved the highest F1-score, indicating its superior ability to balance precision and recall, making it the recommended choice for managing fraudulent activity detection effectively.

MODEL TRAINING AND EVALUATION

- **Hyperparameter Tuning**: Employed `RandomizedSearchCV` to optimize `KNeighborsClassifier` parameters (`n_neighbors`, `weights`, `metric`) using 3-fold cross-validation.
- **Result:**
  - Achieved a high recall (98%) for fraud detection, indicating effective identification of actual fraud cases.
  - Precision for fraud (79%) suggests reliable predictions when fraud is predicted.
  - Overall performance metrics include 78% accuracy and a balanced F1-score of 0.70, showcasing effective classification across both fraud and non-fraud cases.
- This approach ensures the pipeline is optimized for performance, particularly in detecting fraud, and validates its effectiveness with robust evaluation metrics.

PIPELINE

- Built a pipeline using `ColumnTransformer` for preprocessing (one-hot encoding categorical features, scaling numeric features) and integrated a `KNeighborsClassifier` for classification.
- Exported the said pipeline and made predictions loading the same.

Thank you!