Final Report: Fraud Detection Project
Adey Innovations Inc.

1. Project Overview

This project focuses on developing robust machine learning models to detect fraudulent transactions in e-commerce data. The pipeline includes:

- Data preprocessing
- Feature engineering
- Model training and evaluation
- Model explainability using SHAP

2. Data Preprocessing

The following steps were taken to prepare the dataset:

Timestamp Conversion: Transformed raw timestamps into datetime format.
Geolocation Mapping: Mapped IP addresses to countries using geolocation data.
Feature Engineering:
- hour_of_day
- day_of_week
- time_since_signup

Class Imbalance Handling: Applied SMOTE.
Feature Scaling: Used StandardScaler for numerical features.
Categorical Encoding: Applied one-hot encoding to categorical variables.

3. Model Training and Evaluation

- Logistic Regression
- Accuracy: ~66%
- Precision (fraud): 0.16
- Recall (fraud): 0.66
- F1-score (fraud): 0.26
- XGBoost
- Accuracy: ~96%
- Precision (fraud): 0.92
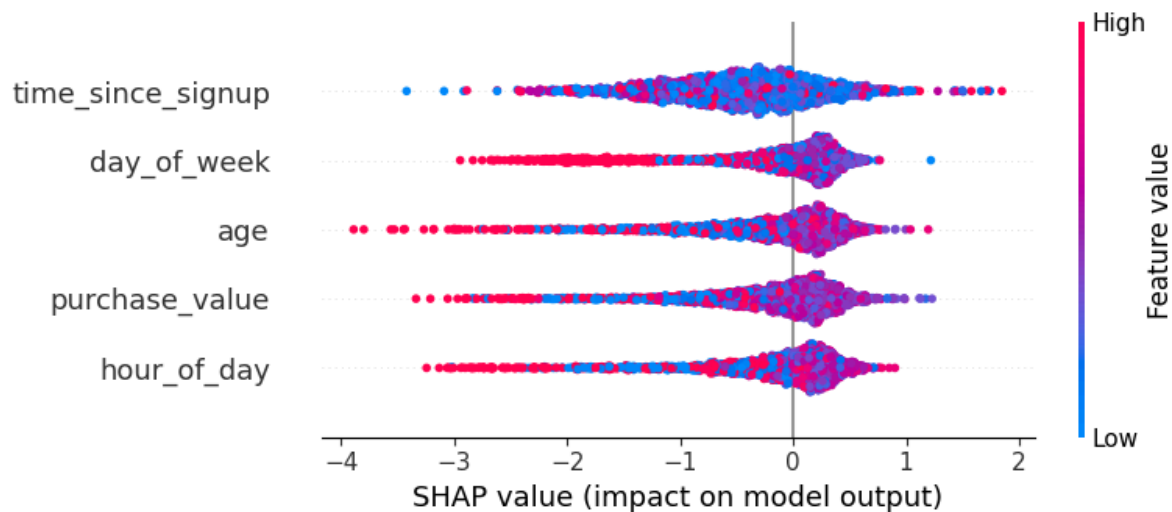- Recall (fraud): 0.57
- F1-score (fraud): 0.70

Conclusion:

XGBoost significantly outperformed Logistic Regression across all metrics and was selected as the final model.

4. Model Explainability with SHAP

SHAP was used to interpret the XGBoost model.

SHAP Summary Plot

Key Insights

Most influential features:

- purchase_value
- time_since_signup
- hour_of_day
- SHAP values provided transparency by showing how each feature contributed to individual predictions.

5. Final Notes

This project delivers a complete fraud detection pipeline with a strong emphasis on:

- Accuracy
- Interpretability
- Business relevance

The integration of SHAP enhances stakeholder trust and provides actionable insights for decision-making.