B5W8: Interim 1 Report - Fraud Detection

1. Overview

This report presents progress on the B5W8 challenge focused on detecting fraud in e-commerce and bank transactions. The tasks completed in Interim 1 include data cleaning, IP-to-country mapping, feature engineering, and exploratory data analysis.

2. Data Cleaning and Preprocessing

- No missing values were found in any of the datasets (Fraud, IP, Credit Card).
- Duplicates were removed from Fraud and CreditCard datasets.
- IP addresses were already in integer format and directly used for merging.

3. IP-to-Country Mapping

IP addresses in the Fraud_Data.csv were enriched with geolocation using the IP ranges provided in IpAddress_to_Country.csv. A custom function was used to find the country for each IP based on its integer value.

4. Feature Engineering

- time_since_signup: Time difference between signup and purchase (in seconds).
- hour_of_day: Hour extracted from purchase timestamp.
- day_of_week: Day of week extracted from purchase timestamp.

5. Exploratory Data Analysis

- The dataset is highly imbalanced: ~90.64% legitimate vs. ~9.36% fraudulent transactions.
- Fraud transactions span all age groups, with more concentration in certain ranges.
- Chrome is the most commonly used browser among both fraud and non-fraud cases.

6. Class Imbalance Strategy

B5W8: Interim 1 Report - Fraud Detection

The fraud class distribution is severely imbalanced. To address this:

- We plan to use SMOTE (Synthetic Minority Over-sampling Technique) or Random Undersampling.

- Resampling will be applied only to the training set to avoid data leakage.

7. Next Steps

- Proceed to model building using Logistic Regression and a tree-based ensemble model (e.g., Random Forest).

- Evaluate models using F1-score, AUC-PR, and confusion matrix.

- Use SHAP to interpret the model predictions.

8. Submission Info

Date: 2025-07-20

Submitted by: Kirubel Gizaw

Challenge: B5W8 - Tenx Platform (10 Academy)