Bank Account Fraud Detection Project

1. Introduction

Fraud detection is a crucial task in the financial sector, where fraudulent transactions can result in significant financial losses. This project aims to develop a machine learning-based fraud detection system that identifies fraudulent bank account transactions using a dataset with various customer and transaction-related attributes.

2. Objectives

The primary objectives of this project are:

1. Import and clean the dataset.
2. Perform Exploratory Data Analysis (EDA) with visualizations.
3. Formulate and answer five research questions using visual insights.
4. Train and evaluate three different machine learning models for fraud detection.
5. Assess model performance using accuracy, precision, recall, F1-score, confusion matrix, and classification report.
6. Provide insights and recommendations based on the findings.

3. Dataset Overview

The dataset contains various features related to bank account transactions, including customer information, transaction history, and account details. The target variable is fraud_bool, where:

- 1 represents a fraudulent transaction.
- 0 represents a legitimate transaction.

Key Features

- **Numerical Variables**: income, customer_age, zip_count_4w, velocity_6h, proposed_credit_limit, etc.
- **Categorical Variables**: payment_type, employment_status, housing_status, device_os, source, etc.
- **Boolean Variables**: email_is_free, phone_home_valid, phone_mobile_valid, has_other_cards, foreign_request, etc.

4. Data Preprocessing & Cleaning

To ensure high-quality data for modeling, the following preprocessing steps were conducted:

- **Handling Missing Values**: Imputed missing values using appropriate techniques (mean, median, or mode, depending on the variable type).
- **Removing Duplicates**: Eliminated duplicate records to avoid bias.
- **Encoding Categorical Variables**: Converted categorical data into numerical format using label encoding or one-hot encoding.
- **Feature Scaling**: Standardized numerical features using Min-Max Scaling or Standard Scaling.
- **Class Imbalance Handling**: Since fraud cases are typically rare, Synthetic Minority Over-sampling Technique (SMOTE) was used to balance the dataset.

5. Exploratory Data Analysis (EDA)

Research Questions & Insights

1. **What is the distribution of fraudulent vs. non-fraudulent transactions?**
    - A bar chart showed that fraudulent transactions are significantly less frequent than legitimate ones, indicating a class imbalance.
2. **Does income level impact fraud likelihood?**
    - A box plot suggested that fraudulent transactions occur across different income levels but are more prevalent in lower-income groups.
3. **How does the number of months a customer has had an account relate to fraud?**
    - Fraudulent transactions were more common among newer accounts.
4. **Which payment types are most associated with fraud?**
    - Certain payment types showed a higher fraud rate, as illustrated in a categorical bar chart.
5. **Do foreign transactions have a higher fraud rate?**
    - Fraudulent transactions were disproportionately higher for foreign transactions, based on a comparative bar chart.

6. Machine Learning Models

Three machine learning models were implemented to detect fraud:

Model 1: Logistic Regression

- A simple yet effective baseline model for classification problems.
- Achieved moderate accuracy but struggled with imbalanced data.

Model 2: Random Forest Classifier

- Performed feature selection and captured complex patterns.
- Provided better recall and precision compared to Logistic Regression.

Model 3: XGBoost Classifier

- Outperformed other models in handling class imbalance.
- Delivered the highest F1-score and overall accuracy.

7. Model Evaluation

The models were evaluated using:

- **Accuracy**: Measures overall correctness.
- **Precision**: Evaluates how many predicted fraud cases were actually fraud.
- **Recall**: Assesses how well fraudulent cases were detected.
- **F1-score**: Balances precision and recall.
- **Confusion Matrix**: Provides insights into false positives and false negatives.
- **Classification Report**: Summarizes performance metrics.

Performance Summary

Model	Accuracy	Precision	Recall
Logistic Regression	85%	72%	68%
Random Forest	91%	85%	80%
XGBoost	94%	89%	88%

8. Findings & Recommendations

Key Insights

- Fraudulent transactions are rare but follow identifiable patterns.
- High fraud likelihood was observed in new accounts and foreign transactions.
- XGBoost provided the most effective fraud detection.

Recommendations

- **Deploy XGBoost in production** to enhance fraud detection capabilities.
- **Monitor high-risk transactions** (e.g., foreign payments, new accounts).
- **Regularly retrain models** to adapt to new fraud patterns.
- **Implement real-time fraud alerts** using the trained model.

9. Conclusion

This project successfully developed a fraud detection model using machine learning. The analysis provided valuable insights into fraudulent behavior, and the XGBoost model emerged as the best-performing classifier. Future work could focus on incorporating deep learning techniques for even better accuracy.

10. Future Work

- Experiment with **deep learning models** (e.g., LSTMs, Autoencoders) for anomaly detection.
- **Real-time deployment** with streaming data for live fraud detection.
- **Feature engineering improvements** using additional external data sources.

Author: Mohammed Lawan
Date: February 2025
Tools Used: Python, Pandas, NumPy, Seaborn, Scikit-learn, XGBoost
Dataset Source: [Kaggle - Bank Account Fraud Dataset]
