Comprehensive report of the assignment Fraud credit card detection:

Credit Card Fraud Detection: Performance Comparison Report

Objective

The objective of this project was to build and evaluate machine learning models to detect fraudulent credit card transactions. The project focused on effectively handling the problem of class imbalance, a common challenge in fraud detection datasets, by applying Random Forest and XGBoost classifiers both with and without SMOTE (Synthetic Minority Over-sampling Technique).

Data Overview

The dataset used for this project was creditcard.csv, sourced from Kaggle. A key characteristic of this dataset is its severe class imbalance, with only **0.1727%** of the transactions being fraudulent. The data contained no missing values and was composed of anonymized features (V1 to V28) as well as Time and Amount features.

Model Performance & Analysis

Three different models were trained and evaluated on a hold-out test set to compare their performance. The key metrics for evaluation were **Precision**, **Recall**, and the **F1-Score**, as overall accuracy is a misleading metric for imbalanced datasets.

Model	Precision	Recall	F1-Score	ROC-AUC
Baseline Random Forest	0.93	0.83	0.88	0.91
Random Forest w/ SMOTE	0.83	0.82	0.82	0.91
XGBoost w/ SMOTE	0.69	**0.85**	0.76	**0.92**

Export to Sheets

- **Baseline Random Forest:** This model, trained on the original imbalanced data, showed high **precision** (0.93), which means it had a very low rate of false alarms. It correctly identified fraudulent transactions in 93% of the cases it flagged as fraud. However, its **recall** was lower (0.83), indicating it missed about 17% of all actual fraudulent transactions.
- **Random Forest with SMOTE:** Surprisingly, applying SMOTE caused the performance of the Random Forest model to drop across all key metrics. This could be due to the synthetic data points confusing the model and introducing noise, making it less effective at making accurate classifications.
- **XGBoost with SMOTE:** This model, despite having the lowest precision (0.69), demonstrated a critical improvement in **recall** (0.85). This means it was better at catching the actual fraudulent transactions, thus minimizing the number of missed fraud cases (false negatives).

Insights & Recommendations

In credit card fraud detection, the cost of a **false negative** (missing a fraudulent transaction) is often much higher than the cost of a **false positive** (flagging a legitimate transaction as fraudulent). Therefore, **recall** is typically the most important metric to optimize.

Based on the results, the **XGBoost model with SMOTE** is the recommended model for deployment. While its precision is lower than the baseline model, its superior **recall** and high **ROC-AUC** score indicate that it is the most effective model at identifying fraudulent activity, which aligns with the primary business objective of minimizing financial losses.

The analysis demonstrates that, for this dataset, **XGBoost's gradient-boosting approach** was more robust to the changes introduced by SMOTE than the bagging-based approach of Random Forest.