Final Capstone Project Report: Fraud Detection Using Machine Learning

1. Introduction

Fraud detection is a critical challenge in the financial services industry. With the rise of digital transactions, identifying fraudulent activity in real time has become increasingly important. This project aims to build a machine learning-based fraud detection system using the Kaggle Credit Card Fraud dataset and deploy it using AWS SageMaker.

2. Problem Statement

The goal is to develop a binary classification model that can accurately identify fraudulent credit card transactions. The dataset is highly imbalanced, with only 0.172% of transactions labeled as fraud. The challenge lies in improving recall while minimizing false positives, ensuring that legitimate transactions are not incorrectly flagged.

3. Data Exploration

- **Dataset**: 284,807 transactions with 30 features (PCA-transformed), including Time, Amount, and Class (target).
- **Fraud Cases**: 492 (0.172%)
- **Missing Values**: None
- **Class Imbalance**: Severe — addressed using scale_pos_weight and stratified splits.
- **Scaling**: Time and Amount were scaled using RobustScaler.

4. Methodology

Workflow:

1. **Data Preprocessing**:
    - Scaled Time and Amount
    - Stratified train/val/test split
    - Saved and uploaded to S3 in both standard and XGBoost-friendly formats
2. **Modeling**:

- Baseline: Logistic Regression using SKLearn estimator
- Advanced: XGBoost with hyperparameter tuning

3. **Evaluation**:
    - Metrics: Precision, Recall, F1-Score, PR-AUC, ROC-AUC
    - Threshold tuning for optimal trade-offs
4. **Deployment**:
    - Batch transform endpoint for inference
    - Model artifacts stored in S3

5. Model Training and Tuning

Baseline Model:

- **Algorithm**: Logistic Regression
- **Framework**: SageMaker SKLearn
- **Performance**: Served as a benchmark

XGBoost Model:

- **Hyperparameters**:
    - objective: binary:logistic
    - eval_metric: aucpr
    - scale_pos_weight: 577
    - num_round: 400
- **Tuning**:
    - max_depth, eta, subsample, colsample_bytree, min_child_weight
    - 12 jobs, 3 parallel

6. Evaluation

Metric	Value
PR-AUC	0.9029
ROC-AUC	0.9820
Precision @ 0.5	0.8775
Recall @ 0.5	0.8775
F1-Score @ 0.5	0.8775
Best F1 Threshold	~0.386
Best Precision ≥ 0.90	0.9111 @ Recall 0.8367

- **Threshold Sweep**: Evaluated multiple thresholds to optimize for business needs (e.g., high recall or high precision).
- **Confusion Matrix**: Used to analyze false positives and false negatives.

7. Deployment

- **Model Artifact**: s3://udacity-fraud-capstone/fraud/outputs/sagemaker-xgboost-250821-1000-002-593cfed0/output/model.tar.gz
- **Batch Transform Output**: s3://udacity-fraud-capstone/fraud/batch-preds/test_nolabel.csv.out
- **Inference**: Batch transform used for scoring test data. Predictions evaluated against true labels.

8. Conclusion

This project successfully built and deployed a high-performing fraud detection model using AWS SageMaker. The XGBoost model achieved strong performance across all key metrics, especially in handling class imbalance. The deployment pipeline is scalable and ready for integration into real-time systems.