upGrad

Capstone Project Report

FRAUD DETECTION MODEL

Prediction of Credit Card fraud (FindDefault)

Presented by
Rajat Kumar

Executive Summary

Problem Statement:

Credit card fraud is a critical issue in the financial sector, causing significant monetary losses and security risks. Traditional fraud detection methods, such as rule-based systems, fail to adapt to evolving fraud patterns, making them inefficient.

Goal of the Study:

The objective of this study is to develop an accurate, scalable, and explainable fraud detection model using Machine Learning while addressing challenges of imbalanced data, feature selection, and real-time deployment.

Key Findings

- The dataset is highly imbalanced **(0.17% fraud cases),** requiring SMOTE for balancing.
- Random Forest emerged as the best model based on precision-recall and AUC-ROC.
- Optimized decision threshold (0.19) improved recall while minimizing false positives.
- The final model is ready for real-time deployment and batch fraud detection.

Deployment Considerations

- Real-Time API Deployment: The model has been successfully deployed as a Flask API, allowing real-time fraud detection via API requests.
- Cloud & On-Premise Readiness: The API is designed to be deployed on cloud platforms (AWS, Render, Heroku) or on-premise financial security systems.
- Fraud Detection API: Users can send transaction data to the /predict endpoint to receive fraud probability and classification.
- Optimized Decision Threshold: The final model applies a threshold of 0.19, improving recall and minimizing false positives.

Introduction

What is Credit Card Fraud?

Credit card fraud occurs when unauthorized transactions are made using a stolen or cloned credit card. It leads to financial losses and compromises consumer trust.

Why is Fraud Detection Important?

- Prevents financial losses.
- Enhances customer trust and security.
- Reduces operational costs related to fraud management.

Challenges in Fraud Detection

- **Class Imbalance**: Fraud cases are rare, making model training difficult.
- **Concept Drift**: Fraud patterns evolve over time.
- **Real-Time Detection Needs:** Fraudulent transactions need to be identified instantly.

Objective of the Study

This study aims to build a scalable, high-performance fraud detection system using Machine Learning models while addressing:

- Class imbalance issues.
- Feature selection & interpretability.
- Real-time fraud detection capabilities.

Dataset & Data Processing

Dataset Overview

- Source Link: Credit Card Fraud Detection Dataset (Provided by upGrad).
- Data Points: **284,807 transactions.**
- Features: **30 anonymized numerical features + Amount + Time**.
- Fraud Cases: Only **0.17%** of transactions are fraudulent.

Preprocessing Steps

- **Feature Scaling:** Standardized Amount and Time for better model performance.
- **Class Balancing:** Applied SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic fraud cases and handle class imbalance.

Feature Engineering

Feature Selection Process

Feature selection helps remove redundant data, reduce model complexity, and improve interpretability.

Approach Used:

- Random Forest Feature Importance to select top 13 features.
- Removed low-impact features that did not contribute significantly to fraud detection.

Feature Importance Analysis

Key Features Driving Fraud Prediction:

- V14, V10, V12, V4, V17, and V3 have the strongest correlation with fraud.
- Higher values of these features indicate a greater likelihood of fraud.
- These insights help financial institutions focus on key risk areas.

Model Selection & Evaluation

Algorithms Considered

1. Logistic Regression (Baseline model, but weak recall)
2. Decision Tree (Good interpretability, prone to overfitting)
3. Random Forest (Best performance & interpretability)
4. XGBoost (Strong but computationally expensive)

Performance Metrics Used

- Accuracy (Overall correctness, but not ideal for imbalanced data)
- Precision & Recall (Key metrics for fraud detection)
- F1-score (Balance between precision & recall)
- AUC-ROC Curve (Evaluates model discrimination power)

Best Model Selection

- **Random Forest achieved best results:**

F1-score	Precision	Recall	AUC-ROC
85%	93%	79%	96%

Why Random Forest?

- Balanced precision & recall for optimal fraud detection.
- Computationally efficient for real-time fraud predictions.
- Feature interpretability supports business decision-making.

Threshold Tuning & Fraud Probability Optimization

Why Default 0.5 Threshold is Not Optimal?

- Standard 0.5 threshold led to high false negatives.
- Optimized Threshold = 0.19 increased recall significantly.

Comparison: Default vs. Optimized Threshold

Threshold	Precision	Recall	F1-Score
0.5	**44%**	**87%**	**58.6%**
0.19 (Optimized)	**23%**	**91%**	**36%**

Fraud Detection Model Deployment

1. **Deployment Pipeline:**
   - Final model trained using **Random Forest** and optimized with **threshold tuning (0.19)**.
   - Model saved as **final_fraud_detection_model.pkl** for seamless loading.
2. **API Development**
   - Built a **Flask API** that accepts transaction data via a POST request.
   - Endpoint /predict allows real-time fraud classification.
3. **Testing & Validation**
   - Successfully tested fraudulent and non-fraudulent transactions using **API calls**.
   - **Example Request:**

```
{
  "features": [-4.3, 2.1, -1.8, 3.0, -2.5, 1.7, -0.9, 2.3, -1.5, 0.6, -0.7, 1.2, 120.5]
}
```

   - **To be Run in Command Prompt:**

```
curl -X POST http://127.0.0.1:5000/predict -H "Content-Type: application/json" -d
"{\"features\": [-4.3, 2.1, -1.8, 3.0, -2.5, 1.7, -0.9, 2.3, -1.5, 0.6, -0.7, 1.2, 120.5]}"
```

- **Expected Output:**

```
{
 "fraud_probability": 0.0313,
 "prediction": "Safe Transaction"
}
```

4. **Real-Time Fraud Prevention & Batch Processing**
   - **Immediate Fraud Screening**: API classifies transactions instantly before payment approval.
   - **Batch Transaction Analysis**: The model supports large-scale fraud detection for financial institutions.

Business Insights & Recommendations

Key Takeaways for Financial Institutions

- Fraud Detection API ensures real-time security against financial fraud.
- Financial Institutions can integrate the model into payment systems for instant fraud screening.
- Automated fraud prevention reduces manual transaction verification efforts, saving operational costs.
- Cloud-ready model enables large-scale fraud monitoring with minimal infrastructure costs.

Conclusion & Future Work

Final Model Selection Summary

- Random Forest emerged as the best model for fraud detection.
- Threshold tuning (0.19) significantly improved fraud recall.
- Final model is ready for real-time & batch processing deployment.

Future Improvements

- Explore Deep Learning (LSTMs, Autoencoders) to detect evolving fraud patterns.
- Implement Anomaly Detection models for adaptive fraud detection.
- Reduce False Positives to avoid unnecessary fraud alerts for customers.

This model is a highly scalable, real-world fraud detection solution, ensuring financial security and trust.

****** End of Report ******