Credit Card Fraud Detection: A Comprehensive Analysis

1. Project Overview

This project addresses the critical challenge of credit card fraud detection through a machine learning approach. The goal is to develop and compare multiple detection models, optimizing their performance through F1-score threshold tuning to achieve a balanced detection of fraudulent transactions.

While in real-world fraud detection, the balance between recall (catching frauds) and precision (avoiding false alarms) would be determined by specific business costs and priorities, this project uses F1-score optimization as a neutral approach since no specific business constraints were provided. This allows for a fair comparison of different models' capabilities in detecting fraud while maintaining reasonable precision.

The project implements and compares multiple detection approaches:

- Supervised learning models (Random Forest, XGBoost, CatBoost)
- Unsupervised learning (Autoencoder-based anomaly detection)
- Various class imbalance handling techniques (SMOTE, ADASYN)

2. Data Description

The dataset consists of credit card transactions with the following characteristics:

- **Total Records**: 284,807 transactions
- **Features**: 30 features (28 PCA-transformed features + Amount + Time)
- **Target Variable**: Binary classification (0: legitimate, 1: fraudulent)
- **Class Distribution**:
    - Legitimate transactions: 99.83%
    - Fraudulent transactions: 0.17% (492 cases)

Key Statistics:

```
Transaction Amount Statistics:
- Mean: $88.35
- Median: $22.00
- Min: $0.00
- Max: $25,691.16

Fraud vs. Legitimate Transaction Amounts:
- Fraud mean: $122.21
- Fraud median: $9.25
- Legitimate mean: $88.29
- Legitimate median: $22.00
```

3. Modeling Approach

3.1 Supervised Learning (Tree-Based Models)

Three powerful tree-based algorithms were implemented:

1. **Random Forest**
    - Ensemble of decision trees
    - Handles non-linear relationships
    - Provides feature importance insights
2. **XGBoost**
    - Gradient boosting framework
    - Optimized for performance and accuracy
    - Handles missing values effectively
3. **CatBoost**
    - Advanced gradient boosting
    - Handles categorical features automatically
    - Reduces overfitting

3.2 Unsupervised Learning (Autoencoder)

- **Architecture**: Symmetric deep neural network
    - Encoder: 64 → 32 → 14 neurons
    - Decoder: 14 → 32 → 64 neurons
- **Training**: Exclusively on legitimate transactions
- **Detection**: Based on reconstruction error
- **Advantage**: Can detect novel fraud patterns

3.3 Class Imbalance Handling

Two oversampling techniques were implemented:

1. **SMOTE** (Synthetic Minority Oversampling Technique)
    - Generates synthetic fraud cases
    - Maintains data distribution characteristics
2. **ADASYN** (Adaptive Synthetic Sampling)
    - Focuses on difficult-to-classify cases
    - Adapts to data distribution

4. Metric Selection

F1-score optimization was chosen as the primary metric because:

- **Balanced Performance**: Combines precision and recall
- **Threshold Optimization**:
  - Models were tuned by adjusting classification thresholds
  - Each model's threshold was optimized to maximize F1-score
  - This provides a balanced approach to fraud detection
- **Practical Impact**:
  - Precision: Proportion of flagged transactions that are actually fraudulent
  - Recall: Proportion of actual fraud cases that are detected

5. Model Performance Summary

Model	Accuracy	F1 Score	Precision	Recall	ROC AUC	PR AUC	Training Time (s)
CatBoost	0.9996	0.8865	0.9425	0.8367	0.9790	0.8743	211.43
XGBoost SMOTE	0.9996	0.8830	0.9222	0.8469	0.9760	0.8795	211.20
XGBoost	0.9996	0.8827	0.9753	0.8061	0.9778	0.8848	182.33
XGBoost ADASYN	0.9996	0.8634	0.9294	0.8061	0.9829	0.8675	217.76
Random Forest	0.9995	0.8632	0.8913	0.8367	0.9519	0.8614	214.18
Autoencoder	0.9921	0.7585	0.7920	0.7276	0.9448	0.7644	20.05

6. Visualizations

6.1 t-SNE Visualization

- Shows clear separation between legitimate and fraudulent transactions
- Helps understand the data distribution in reduced dimensions
- Demonstrates the effectiveness of feature engineering

6.2 Reconstruction Error Distribution

- Autoencoder's reconstruction error histogram shows:
  - Clear separation between normal and fraudulent transactions
  - Optimal threshold for classification
  - Effectiveness of anomaly detection approach

7. Final Recommendation

7.1 Primary Model Selection

- **CatBoost** is recommended as the primary model due to:
  - Highest F1-score (0.8865)
  - Best balance of precision (0.9425) and recall (0.8367)
  - Strong ROC AUC (0.9790) and PR AUC (0.8743)

7.2 Implementation Strategy

1. **Deploy CatBoost as primary detector**
   - Use optimized threshold for F1-score
   - Implement real-time monitoring
   - Regular model retraining
2. **Complementary Systems**
   - Use autoencoder for novel fraud pattern detection
   - Implement ensemble approach for high-risk transactions
   - Maintain separate models for different transaction segments

This comprehensive approach provides a robust fraud detection system that balances detection accuracy with operational efficiency, while maintaining the flexibility to adapt to evolving fraud patterns.