Statistical Learning Project: Fraud Detection

This report explores using supervised and unsupervised learning techniques for credit card fraud detection. The analysis is based on a highly imbalanced dataset of credit card transactions, with only 0.172% being fraudulent. The project examines both XGBoost for supervised classification and Principal Component Analysis (PCA) for unsupervised dimensionality reduction and visualization.

S by Salima Tankibayeva

Dataset Overview

The Credit Card Fraud Detection dataset contains 284,807 transactions from European cardholders over two days in September 2013. Only 492 transactions are fraudulent, creating a significant class imbalance. Most features (V1-V28) are PCA-transformed for privacy, while Time and Amount remain untransformed. The target variable "Class" indicates fraudulent (1) or legitimate (0) transactions.

1

Total Transactions

284,807

2

Fraudulent Transactions

492 (0.172%)

3

Features

28 PCA-transformed (V1-V28), Time, Amount

Supervised Learning: XGBoost Model

The XGBoost algorithm was used for supervised learning, with data preprocessing including scaling and addressing class imbalance. The model was trained using parameters such as binary:logistic objective, logloss evaluation metric, and scale_pos_weight to handle imbalance. Cross-validation was employed with early stopping to prevent overfitting.

Preprocessing

Scaling and class imbalance handling

Algorithm

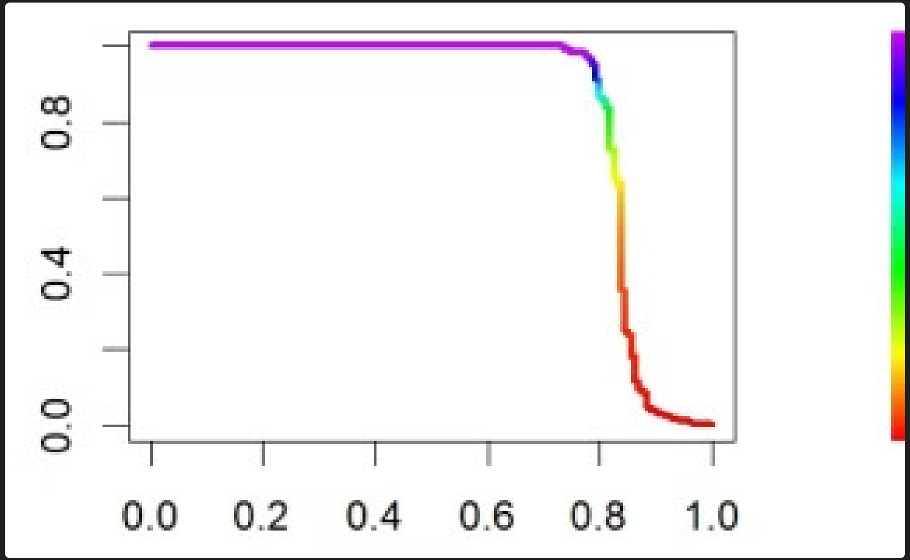
XGBoost with binary:logistic objective

Training

Cross-validation with early stopping

XGBoost Model Evaluation

The XGBoost model was evaluated using a confusion matrix and Precision-Recall curve. The Area Under the Precision-Recall Curve (AUPRC) was calculated to be 0.8374, indicating good performance in balancing precision and recall. This metric is particularly important given the significant class imbalance in the dataset.



Precision-Recall Curve

Visualization of model performance across different thresholds

Feature Importance in XGBoost

Feature importance analysis was conducted to understand which variables contributed most to the model's predictions. This provides insights into the key factors driving fraud detection in the XGBoost model.

1

Identify Top Features

Analyze XGBoost model to determine most influential variables

2

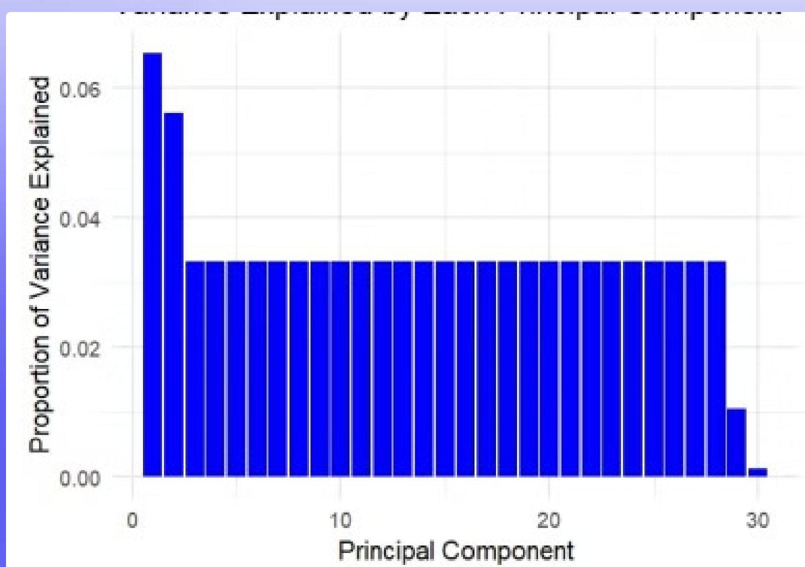
Visualize Importance

Create bar chart or similar visualization of feature importance scores

3

Interpret Results

Understand which features are most crucial for fraud detection



Unsupervised Learning: PCA Analysis

Principal Component Analysis (PCA) was applied as an unsupervised learning technique for dimensionality reduction and visualization. Data preprocessing included handling missing data, removing constant and zero-variance features, and scaling. The explained variance for each principal component was analyzed to assess information retention.

1. **Preprocessing**

   Handle missing data, remove constant features, scale

2. **Apply PCA**

   Reduce dimensionality while retaining variance

3. **Analyze Results**

   Examine explained variance and visualize data

PCA Results and Limitations

PCA results showed limited effectiveness for fraud detection. The first two principal components explained only 12.1% of total variance, with 22 components needed to retain 78.8% variance. Visualization of the first two components failed to separate fraudulent from non-fraudulent transactions. This highlights PCA's limitations in handling highly imbalanced datasets for fraud detection when used in isolation.

Explained Variance

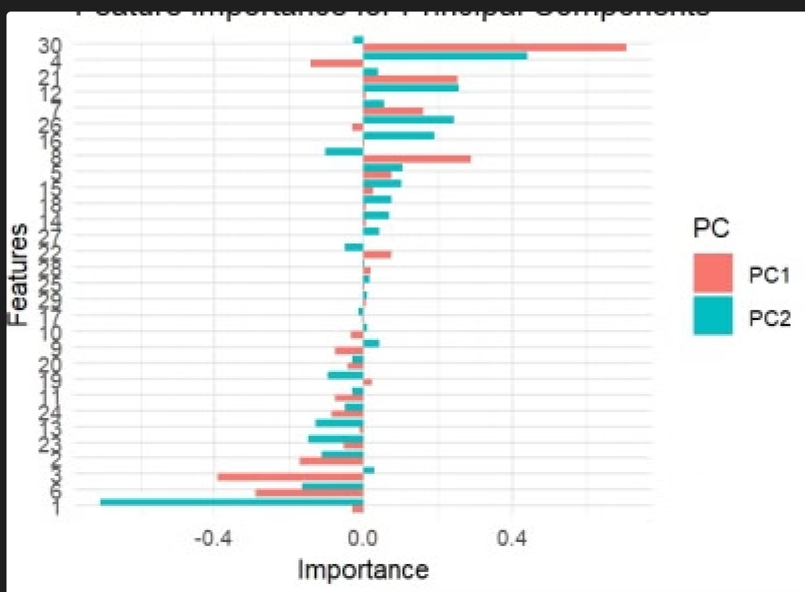
PC1: 6.5%, PC2: 5.6%, 22 PCs for 78.8% total

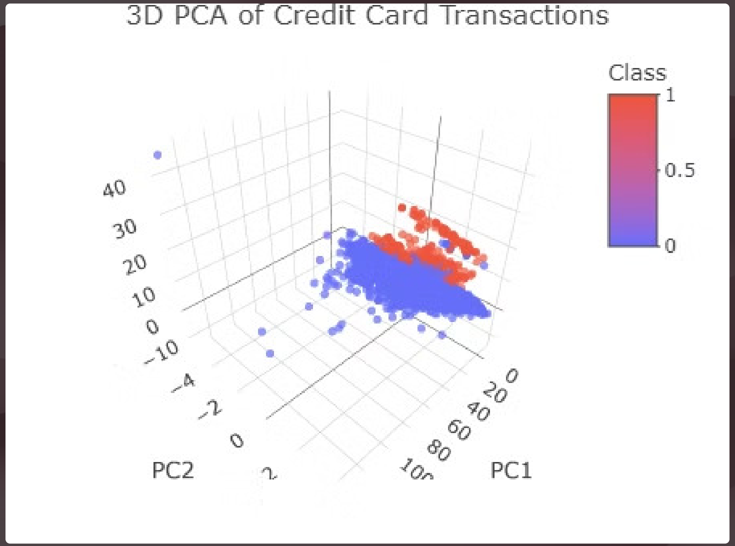
Visualization

No clear separation between classes in PCA plot

Limitation

Ineffective for fraud detection in imbalanced data




3D PCA of Credit Card Transactions
Class
1
0.5
0
40
30
20
10
0
−10
−4
−2
0
PC2
0
20
40
60
80
PC1

Conclusion and Future Work

The analysis demonstrated the effectiveness of supervised learning (XGBoost) for fraud detection in imbalanced datasets, while highlighting the limitations of unsupervised PCA. Future work should focus on combining supervised and unsupervised methods, incorporating advanced anomaly detection techniques, and improving methods for handling class imbalance. This hybrid approach could leverage the strengths of both methods to create more robust fraud detection systems.

Hybrid Approach

Combine supervised and unsupervised methods

Anomaly Detection

Incorporate advanced techniques

Class Imbalance

Improve handling of imbalanced data