An Analytic Approach of Bank Account Fraudulent Prediction

Chin-Hsien Tsai, Yanhui Li, Mina Liu, Cheng Feng, Xiaofan Jiao, Xin Chen

Introduction

In the realm of financial security, detecting and preventing fraudulent activities in bank account transactions is paramount. Leveraging the BAF dataset for its advanced privacy measures, our study aims to improve fraud detection accuracy in bank transactions by addressing gaps in current models. We utilized Python for both modeling and interactive visualizations, enhancing our understanding and analysis of financial fraud complexities.

Our Data

The Bank Account Fraud (BAF) suite of datasets was used. It has been published at NeurIPS 2022 and it comprises a total of 6 different synthetic bank account fraud tabular datasets. The 21 variables were selected for analysis.

Proposed Methods

Data Cleaning

- Conducted data cleansing and correlation analysis to reduce multicollinearity and used Principal Component Analysis to select variables that enhance fraud detection model performance.

Classification Methods

- Base Method
- Transformation Method - Box-Cox, logarithmic
- Polynomial Method

Clustering Methods

- Focus on Numerical Data
- PCA Utilization

Algorithm

- **Logistic Regression** - Effective for binary outcomes in fraud detection.
- **Random Forest** - Robust and manages complex data without overfitting.
- **Naive Bayes** - Fast and efficient for large datasets, despite its simplicity.
- **Ensemble Learning** - Combines multiple models to improve accuracy and reduce errors.
- **K-means Clustering** - Applied to datasets using the Elbow method to optimize cluster numbers.
- **Gaussian Mixture Model (GMM)** - Evaluated using BIC to determine optimal cluster configurations for diverse shapes and sizes.

Innovation

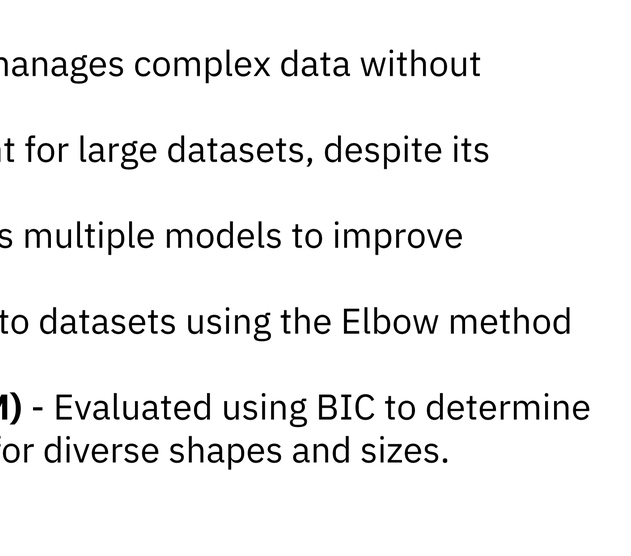
Diverse Model Integration - Combines logistic regression, random forest, and Naive Bayes for robust fraud detection.

Advanced Ensemble Techniques - Enhances accuracy and reliability by merging strengths of multiple models.

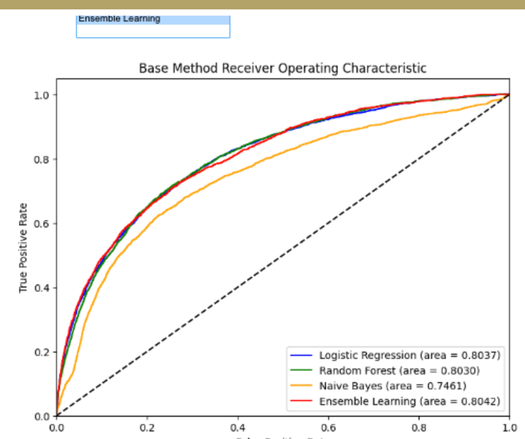
Sophisticated Clustering - Used K-means and Gaussian Mixture Models for nuanced fraud pattern identification.

Visualizaiton

Interactive Confusion Matrix: Utilized confusion matrices with interactive features in numpy and matplotlib.pyplot, enhanced by ipywidgets, to effectively visualize and evaluate the strengths and weaknesses of models like Logistic Regression in fraud detection.

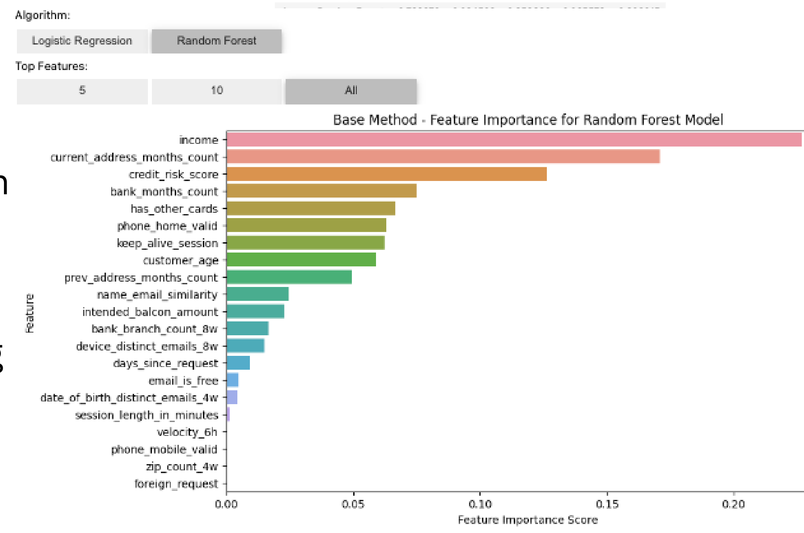


Interactive ROC AUC Plots: Our research employs plotly.express and plotly.graph_objects, enhanced by ipywidgets, to enable users to interactively compare the performance of classification models via color-coded ROC curves with direct AUC annotations, alongside a comprehensive table of key metrics like accuracy and precision.

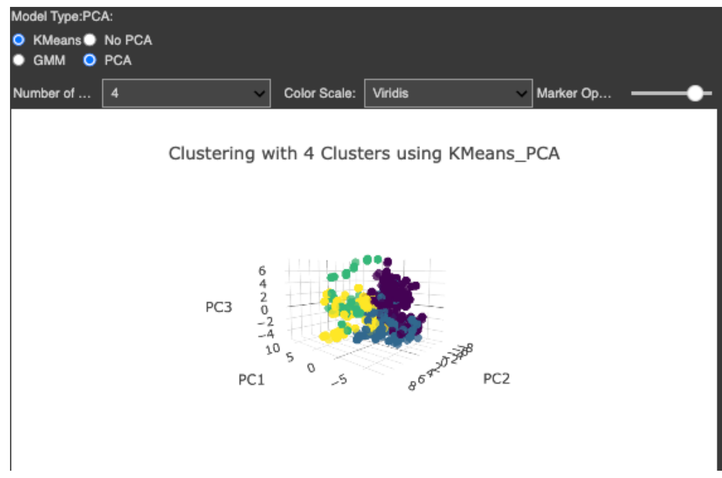


	Model	Accuracy	Precision	Recall	F1 Score	AUC
0	Logistic Regression	0.744640	0.030084	0.707052	0.057712	0.803747

Interactive Feature Importance Visualization: utilized pandas, seaborn, matplotlib for visualizations, focusing on Logistic Regression and Random Forest to highlight top predictive features via interactive bar plots, with ipywidgets enabling user-driven customization for an immediate visual assessment of feature significance in fraud detection.



Interactive 3D Cluster Analysis: Used Plotly to create dynamic 3D plots for K-means and GMM clustering, allowing users to interactively adjust views, cluster numbers, and data types, which aids in model validation and helps determine the optimal clustering configurations essential for identifying fraudulent patterns.



Evaluation

- **Logistic Regression** demonstrated solid fraud detection with an optimal balance of accuracy and recall across different methods, achieving up to 76.83% accuracy and 72.20% recall.
- **Random Forest** maintained high accuracy up to 79.86% and a consistent recall around 68.49%, proving effective in managing complex data relationships in fraud detection.
- **Naive Bayes** excelled in recall, especially in the Polynomial Method, reaching 78.75%, making it highly effective for identifying fraud despite lower accuracy.
- **Ensemble Learning** combined strengths of various models to achieve robust recall rates up to 74.01%, enhancing overall fraud detection performance.
- **K-Means** without PCA achieved the highest Average Silhouette Score of 0.60838, indicating well-separated clusters, which was superior to K-Means with PCA.
- **GMM** with PCA showed a lower score of 0.339224, indicating less distinct clustering, while GMM without PCA improved significantly to 0.44215, demonstrating better-defined cluster structures.

Conclusion

- Developed machine learning framework for fraud detection, incorporating data preprocessing, PCA, and interactive visuals like ROC and 3D plots.
- Recommend future studies using live data or unsupervised learning to improve and adapt fraud detection models.

Summarize Table (Base, Transformation, Polynomial)

Method	Logistic Regression				Random Forest				Naive Bayes				Ensemble			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Base	0.7446	0.0301	0.7071	0.0577	0.7924	0.0345	0.6587	0.0656	0.6716	0.0237	0.7125	0.0458	0.7484	0.0308	0.7134	0.059
Method1 - Transformation	0.7528	0.031	0.7048	0.0593	0.7986	0.0354	0.6551	0.0671	0.682	0.0241	0.7034	0.0466	0.7576	0.0316	0.7061	0.0605