Credit Card Fraud Detection Using ML Techniques & Comparing

Diren Boran Sezen
21946553

İlkim İclal Aydoğan
21992814

Hasan Çağrı Sarıkaya
2200356852

BBM-480 Design Project
Supervisor : Dilmurod Vahabdjanov

Introduction

The Problem: Credit card fraud is a pervasive and costly problem, causing billions of dollars in losses annually. Traditional rule-based systems struggle to keep pace with the evolving tactics of fraudsters.

Our Goal: To develop more accurate and adaptable fraud detection models using machine learning and anomaly detection techniques.

Our Approach: We leverage diverse datasets, feature engineering, and geographic enrichment to enhance model performance and capture complex fraud patterns.

Datasets

Real-World Transactions:
Source: Kaggle
Size: 284,807 transactions
Features: 30 numerical features (PCA transformed)
Challenge: Highly imbalanced (few fraudulent transactions)
Simulated Transactions (IBM TabFormer):
Source: IBM Research
Size: 24,386,900 transactions
Features: User, card, merchant, transaction details, errors
Advantage: Rich feature set for comprehensive modeling

Evaluation Metrics:

- 5-fold cross-validation
- Metrics: Precision, Recall, F1-score, AUC-ROC

Neural Network

Artificial Neural Network (ANN):

- Three fully connected layers
- Tanh activation functions.
- Trained using Adam optimizer and cross-entropy loss

Models

Supervised Learning Algorithms

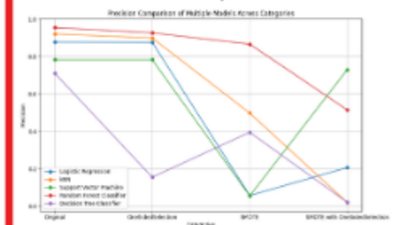
- Logistic Regression: Linear model for binary classification, suitable for predicting whether a transaction is fraudulent or not.
- k-Nearest Neighbors (KNN): A non-parametric method that classifies transactions based on their similarity to neighboring data points in the feature space.
- Support Vector Machine (SVM): Model that finds the optimal hyperplane to separate fraudulent and legitimate transactions.
- Random Forest (RF): An ensemble learning method that combines multiple decision trees to improve prediction accuracy and robustness
- Decision Tree (DT): A tree-like model that makes decisions based on a series of hierarchical rules learned from the data.

ANOMALY DETECTION

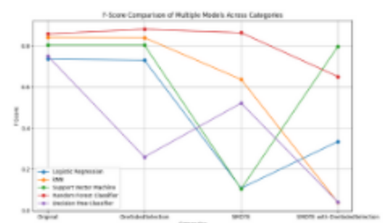
- One-Class SVM: Learns a profile of normal transactions and identifies outliers that deviate significantly from this profile.
- Isolation Forest: Isolates anomalies by randomly partitioning the feature space and identifying points that are easier to isolate.

Results

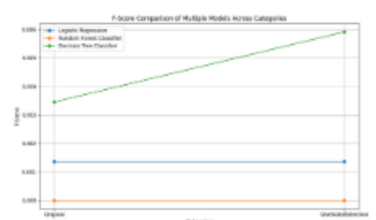
Precision Comparison



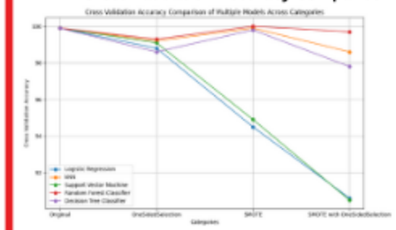
F-Score comparison



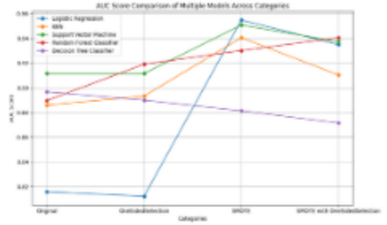
F-Score comparison (Tabformer)



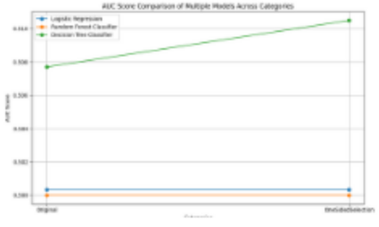
Cross Validation Accuracy Comparison



AUC Score Comparison



AUC Score Comparison (Tabformer)



One Class Support Vector Machine

Classification Report:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	85307
1	0.07	0.64	0.13	136
accuracy			0.99	85443
macro avg	0.54	0.81	0.56	85443
weighted avg	1.00	0.99	0.99	85443

With Customized Threshold

Classification Report:

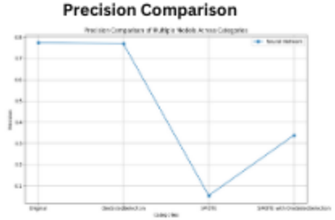
	precision	recall	f1-score	support
0	1.00	0.97	0.99	85307
1	0.04	0.77	0.08	136
accuracy			0.97	85443
macro avg	0.52	0.87	0.53	85443
weighted avg	1.00	0.97	0.98	85443

Isolation Forest

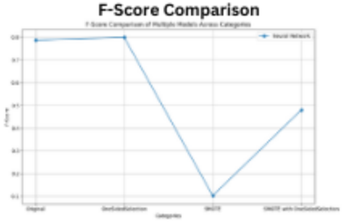
Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	85307
1	0.22	0.51	0.31	136
accuracy			1.00	85443
macro avg	0.61	0.75	0.65	85443
weighted avg	1.00	1.00	1.00	85443

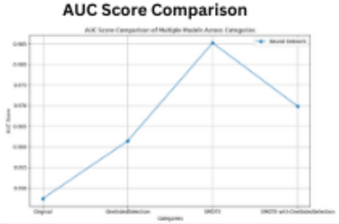
Precision Comparison



F-Score Comparison



AUC Score Comparison



Classification Report for tabformer data

Accuracy on test set: 0.93
Classification Report:

	precision	recall	f1-score	support
0	1.00	0.93	0.96	3653647
1	0.01	0.50	0.02	4388
accuracy			0.93	3658035
macro avg	0.50	0.72	0.49	3658035
weighted avg	1.00	0.93	0.96	3658035

AUC Score: 0.80