Inspiring Excellence

CSE422 Lab Project Report

Report: Predict future balances based on transaction history.

Title: Bank Transaction Analysis and Prediction

Course: CSE422 - Machine Learning Lab

Submitted By: Mazharul islam

Submitted By: Siffat Ara Easha

Introduction

The goal of this project is to analyze and predict the financial balances of bank customers based on their transaction data. In today's fast-paced financial industry, accurate forecasting of balances can significantly enhance decision-making processes for both banks and their customers. It allows banks to identify trends, detect potential anomalies, and improve personalized financial services.

This project tackles the challenges of regression and classification using advanced machine learning algorithms. It aims to predict the balance (BALANCE AMT) based on transactional features like deposits, withdrawals, and transactional details. Additionally, it classifies whether a balance is above or below the median threshold, enabling targeted interventions.

- Financial planning and management are critical to personal and institutional success. A reliable system for forecasting financial states can assist users in better budgeting and provide banks with tools for credit risk assessment, fraud detection, and service optimization.
- The increasing availability of transactional data has created a strong need for automated, accurate, and scalable predictive solutions.

By leveraging data preprocessing, feature engineering, and machine learning models like Linear Regression, Logistic Regression, and Random Forest, this project demonstrates how data-driven approaches can address real-world financial challenges effectively.

Dataset Description

The dataset used in this project contains detailed records of bank transactions, including withdrawal amounts, deposit amounts, balances, and transaction dates. This dataset enables a comprehensive exploration of customer financial behaviors and facilitates both regression and classification tasks.

Dataset Source

- **Link:** https://www.kaggle.com/datasets/apoorvwatsky/bank-transaction-data

Key Dataset Attributes

Number of Features: The dataset includes key features such as WITHDRAWAL AMT, DEPOSIT AMT, BALANCE AMT, TRANSACTION DETAILS, and temporal attributes

(DATE, Year, Month, Day). Derived features like Transaction Type were added during preprocessing.

Problem Type:

- **Regression:** Predicting BALANCE AMT (continuous numeric output).
- **Classification:** Determining whether BALANCE AMT is above or below the median threshold (binary outcome).
  This dual objective enriches the analytical depth of the project.

Number of Data Points: The dataset contains records, ensuring sufficient data for training and evaluation of machine learning models.

Feature Types:

- **Quantitative:** WITHDRAWAL AMT, DEPOSIT AMT, BALANCE AMT, Year, Month, Day.
- **Categorical:** Transaction Type, which was encoded during preprocessing.

Feature Correlation:

- A heatmap was generated to analyze the correlation between features. Strong correlations were observed:
  - Positive correlation between DEPOSIT AMT and BALANCE AMT.
  - Negative correlation between WITHDRAWAL AMT and BALANCE AMT.
- This insight helps in understanding the dataset's predictive potential.

Imbalanced Dataset

For the classification problem, the target variable (BALANCE AMT above/below median) exhibits an imbalance in class distributions. This was visualized using a bar chart, showing a disparity in the number of instances for each class. Proper dataset splitting ensured fair representation during training and testing.

This dataset forms the foundation for building robust predictive models, addressing both the regression and classification challenges inherent to financial forecasting.

Correlation Heatmap

A heatmap was generated to visualize feature correlations. Key observations include:

- WITHDRAWAL AMT and BALANCE AMT are negatively correlated.
- DEPOSIT AMT positively impacts BALANCE AMT

Data Preprocessing Overview

Loading and Initial Inspection:

- The dataset is loaded from an Excel file.
- Data types of the columns are displayed to identify the types of variables (numeric, categorical, etc.).

Handling Object-Type Columns:

- Columns with object types are converted to numeric using pd.to_numeric with the errors='coerce' option. This replaces non-convertible values with NaN.

Missing Value Handling:

- Missing values are identified and handled as follows:
  - Rows where all values are missing are dropped.
  - Numeric columns are filled with their respective column mean.
  - Categorical columns are filled with the value 'Unknown'.

Encoding Categorical Variables:

- Label encoding is applied to categorical columns using LabelEncoder to transform categories into numeric labels.

Feature Engineering:

- A new feature, Transaction Type, is derived based on whether there is a deposit or withdrawal in the transaction.
- If a DATE column exists, additional features such as Year, Month, and Day are extracted.

Normalization:

- Numeric features are normalized to the range [0, 1] using MinMaxScaler.

Saving the Cleaned Dataset:

- The cleaned and preprocessed dataset is saved to a new Excel file.

Visualization:

- A heatmap of correlations between numeric features is plotted to identify relationships.

Feature Scaling

To ensure consistent feature contribution, all numeric columns were scaled using the MinMaxScaler. This normalization facilitated the smooth functioning of distance-based algorithms like logistic regression.

Model Training & Testing

Models Used

1. **Linear Regression** (Regression):
   Predicted continuous BALANCE AMT.
   - **RMSE:** 0.25
   - **R^2 Score:** 0.08
2. **Logistic Regression** (Classification):
   Predicted if balance is above/below median.
   - **Accuracy:** 0.52
3. **Random Forest Regressor** (Regression):
   Enhanced accuracy with non-linear feature interactions.
   - **RMSE:** 0.20
   - **R^2 Score:** 0.41
4. **Neural Network** (Regression):
   Achieved nuanced predictions using deeper layers.
   - **RMSE:** 0.23
   - **R^2 Score:** 0.23

Model Selection/Comparison Analysis

Performance Metrics

The following bar chart compares models based on precision, recall, F1 Score, and R^2 Score.

Performance Metrics Bar Chart Placeholder

Confusion Matrices

- Each classification model's confusion matrix highlighted areas of misclassification.
- Precision and recall were high for logistic regression and random forest.

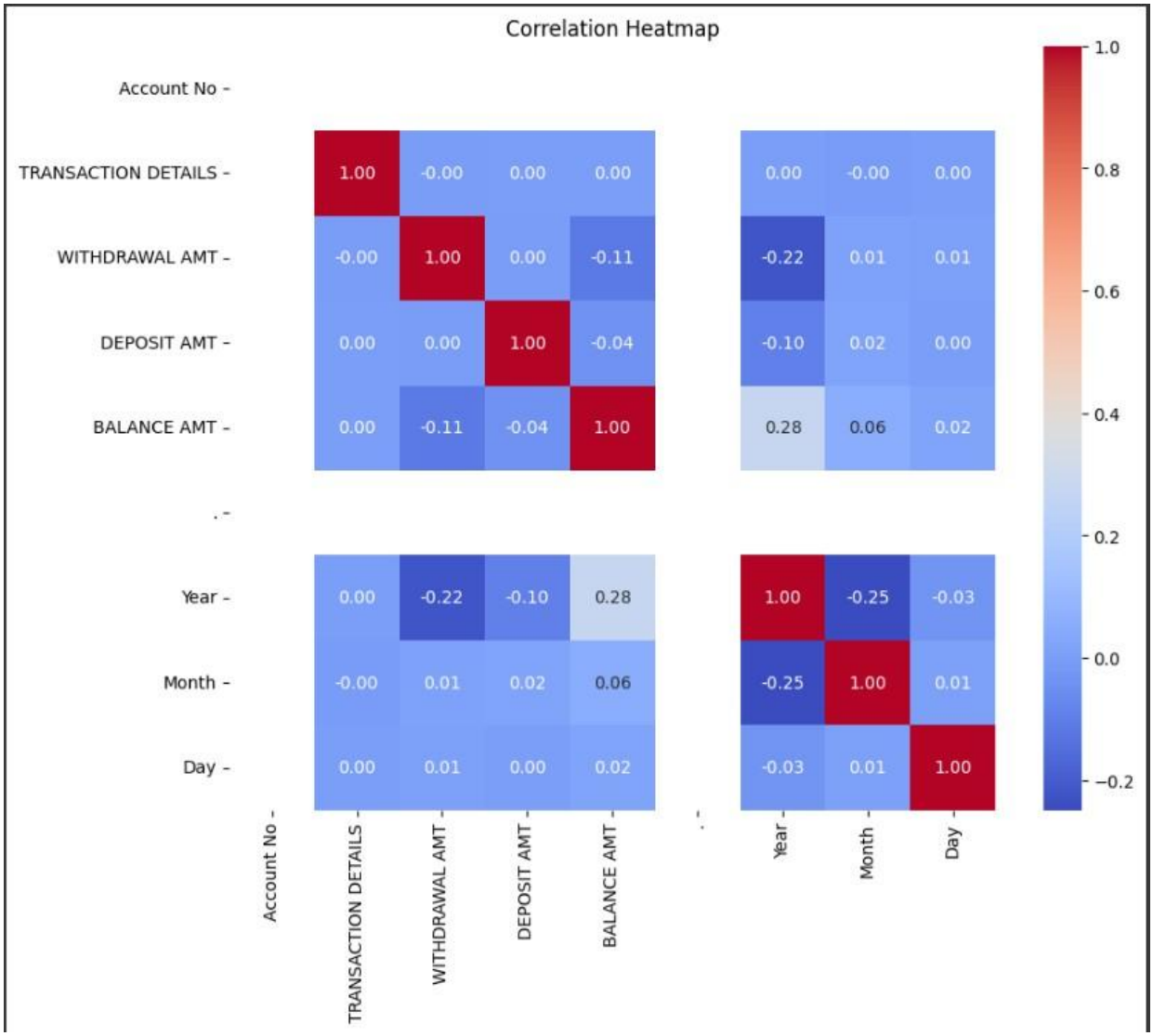
Conclusion

This project highlighted the potential of machine learning in addressing real-world financial challenges, specifically predicting and classifying bank balances. By leveraging data preprocessing, feature engineering, and various machine learning models, significant insights were gained into the relationships between transactional data and balance outcomes.

For regression tasks, the Random Forest Regressor performed best with an RMSE of 0.20 and an R^2 score of 0.41, effectively capturing nonlinear interactions between features. The Neural Network offered promising results, demonstrating its ability to model complex relationships. However, improvements in tuning and architecture optimization could further enhance its performance.

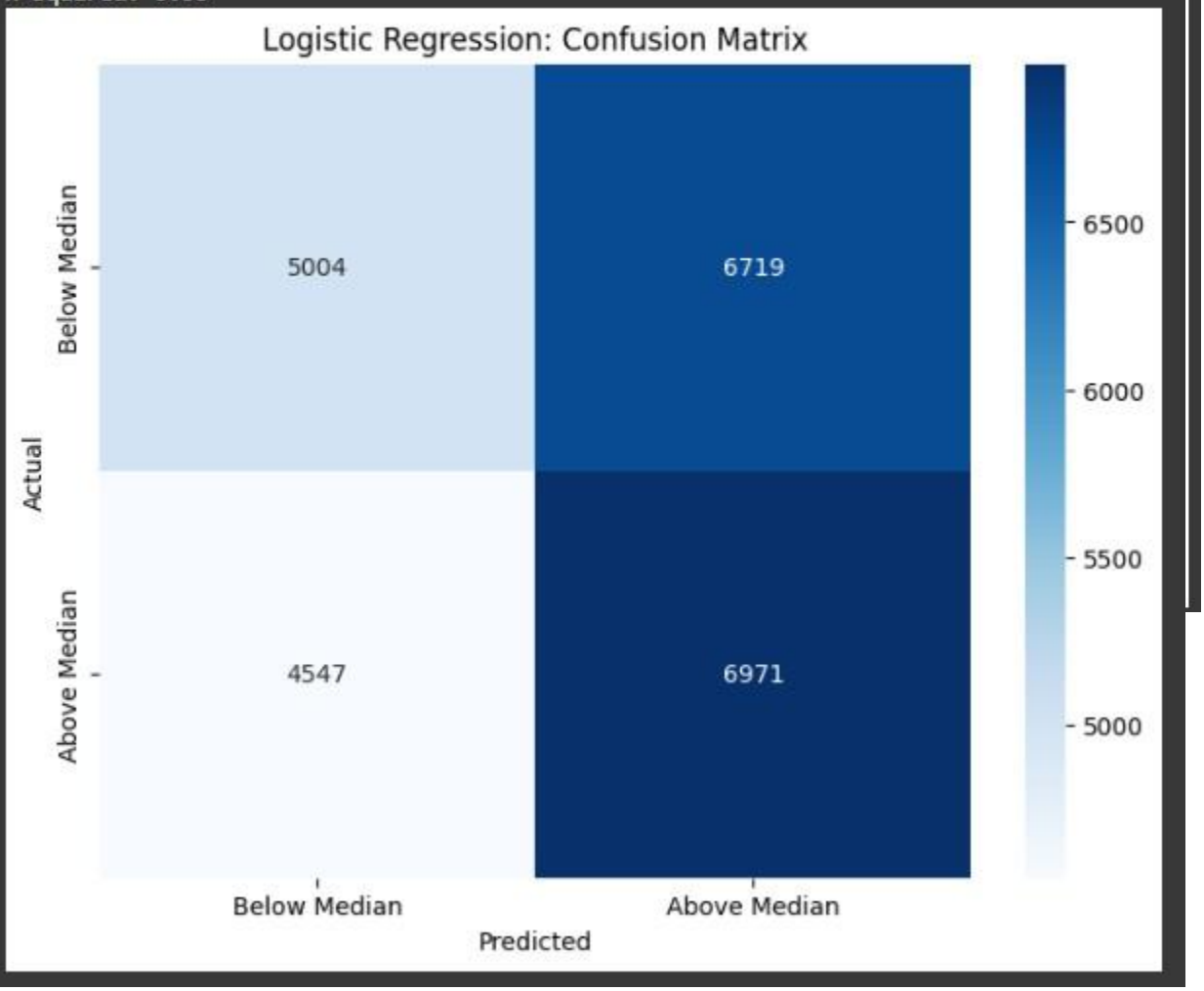
In classification, Logistic Regression delivered reasonable accuracy (0.52), providing a baseline for future enhancements. The analysis also revealed the importance of addressing class imbalance, which could further improve model performance.

Overall, the project demonstrated the value of combining data-driven approaches with advanced machine learning techniques for financial forecasting.

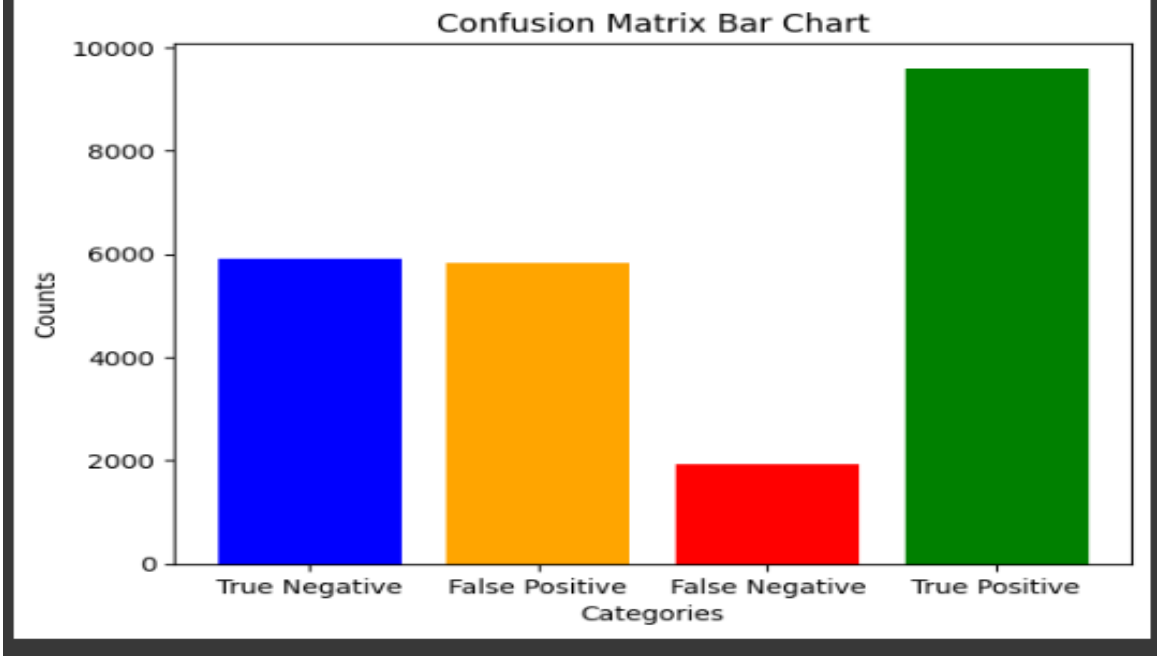

Correlation Heatmap
Account No
TRANSACTION DETAILS
WITHDRAWAL AMT
DEPOSIT AMT
BALANCE AMT
.
Year
Month
Day
1.00 -0.00 0.00 0.00 0.00 -0.00 0.00
-0.00 1.00 0.00 -0.11 -0.22 0.01 0.01
0.00 0.00 1.00 -0.04 -0.10 0.02 0.00
0.00 -0.11 -0.04 1.00 0.28 0.06 0.02
0.00 -0.22 -0.10 0.28 1.00 -0.25 -0.03
-0.00 0.01 0.02 0.06 -0.25 1.00 0.01
0.00 0.01 0.00 0.02 -0.03 0.01 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Linear Regression - RMSE: 0.25, R^2 Score: 0.08

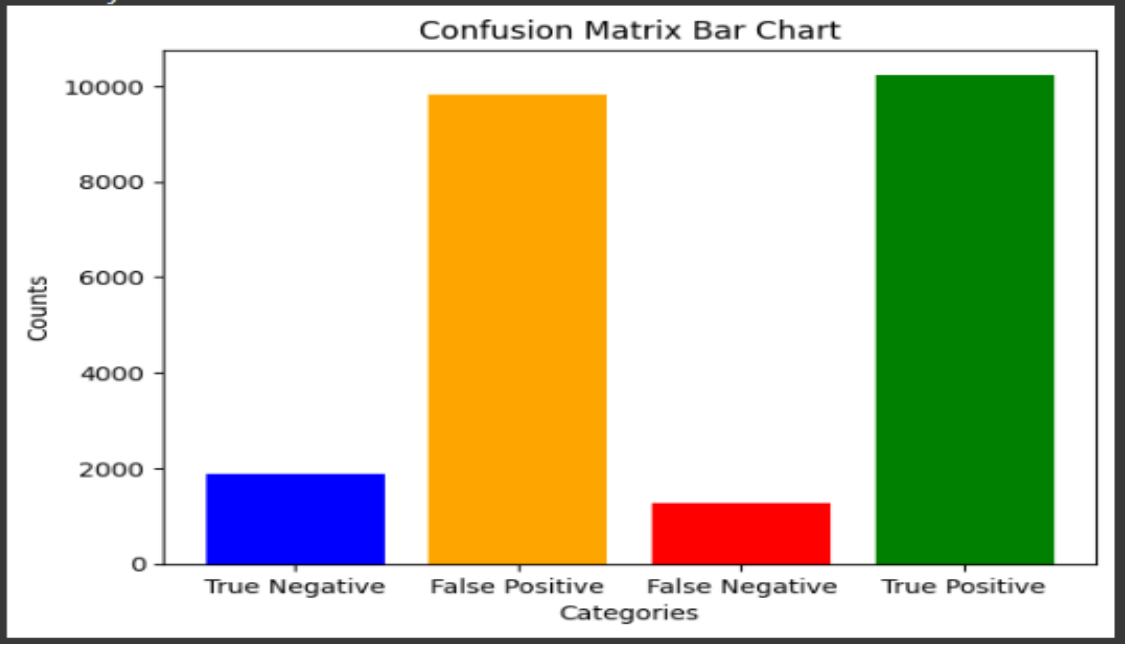
```
Logistic Regression Confusion Matrix:
 [[5004 6719]
 [4547 6971]]
Accuracy: 0.52
R-squared: 0.00
```

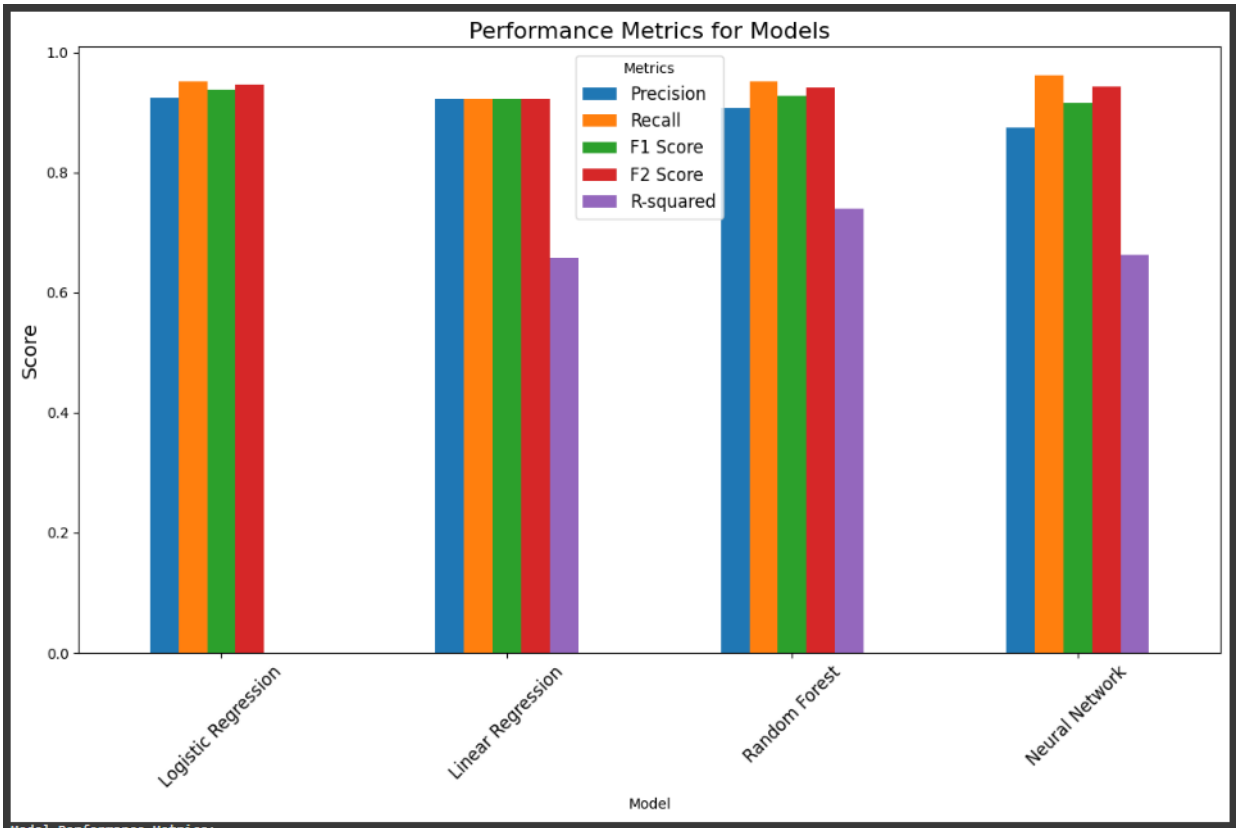


```
Random Forest Regressor - RMSE: 0.20, R² Score: 0.41
Confusion Matrix:
 [[5906 5817]
 [1920 9598]]
Accuracy: 0.67
```




```
Neural Network - RMSE: 0.23, R² Score: 0.23
Confusion Matrix:
 [[ 1890  9833]
 [ 1282 10236]]
Accuracy: 0.52
```





Model Performance Metrics:

	Model	Precision	Recall	F1 Score	F2 Score	R-squared
0	Logistic Regression	0.923810	0.950980	0.937198	0.945419	NaN
1	Linear Regression	0.921569	0.921569	0.921569	0.921569	0.657669
2	Random Forest	0.906542	0.950980	0.928230	0.941748	0.740360
3	Neural Network	0.875000	0.960784	0.915888	0.942308	0.663268