Risk Tier Classification Based on User Behavior

Introduction

This project classifies users into Low, Medium, and High Risk tiers of Being Rigged Based on their Financial Behavior using a synthetic dataset and machine learning models.In this project, we developed a classification-based machine learning framework to seg- ment users into risk tiers Low, Medium, and High based on synthetic behavioral data. Two models were implemented and evaluated viz Logistic Regression and Random Forest Classifier.

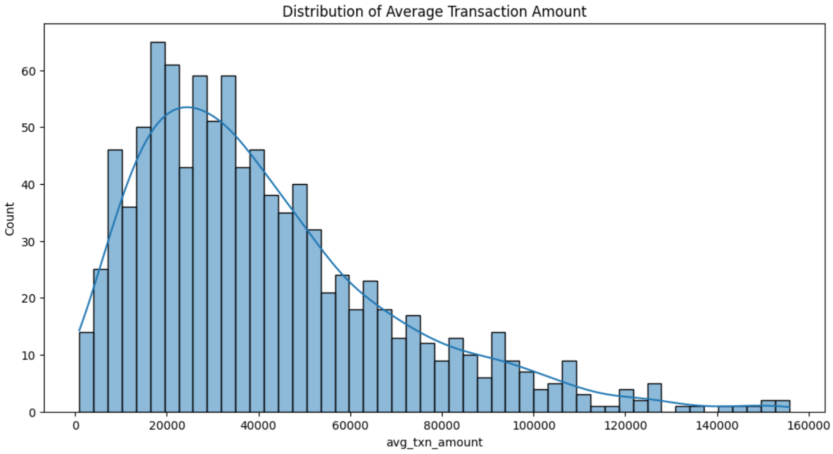
Dataset

Synthetic data set of 1000 users

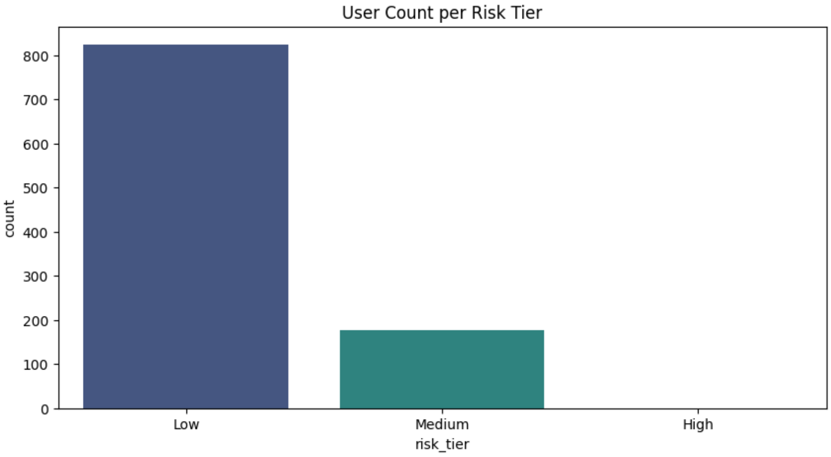
- features: Average transaction amount, Transaction frequency per day, Night transaction ratio, Geographical diversity score, Device changes in last 30 days, Foreign transaction count, Return transaction ratio, Linked accounts count, Risk tier code
- The target variable `risk_tier` has 3 classes: *Low*, *Medium*, and *High* risk.

3. Exploratory Data Analysis

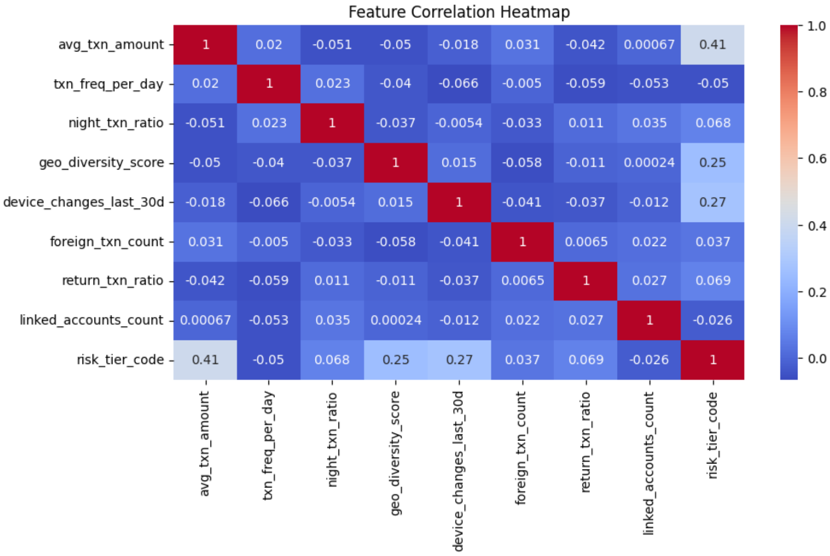
- Distribution of average transaction amounts showed a right-skewed behavior.



- Risk tier distribution was moderately imbalanced.



- A correlation heatmap indicated relationships between credit utilization, delayed payments, and risk.



Methodology

Random Forest classifier and Logistic Regression classifier trained with 80/20 split. Both Random Forest and Logistic Regression classifiers were evaluated. GridSearchCV was applied for each model to identify optimal hyperparameters using 5-fold cross-validation.Feature scaling (StandardScaler) was applied only to logistic regression, as tree-based models like Random Forest do not require scaling. Each model was evaluated on the test dataset using:

Accuracy, F1-Score (Macro Average), Classification Report (Precision, Recall, F1-score per class), Confusion Matrix (Visualized using Seaborn heatmaps).

A side-by-side bar chart was plotted to compare the performance of both models on accuracy and F1-score.

Results

Model Performance Comparison

The classification performance of Logistic Regression and Random Forest was evaluated on the test set using accuracy, macro-averaged F1-score, confusion matrices, and classification reports.

Table 1: Classification Report - Logistic Regression(with GridSearchCV)

Class	Precision	Recall	F1-score	Support
Low Risk	0.94	0.96	0.95	171
Medium Risk	0.73	0.66	0.69	29
High Risk	0	0	0	0
Accuracy		0.92		
Macro Avg	0.84	0.81	0.82	200
Weighted Avg	0.91	0.92	0.91	200

Table 2: Classification Report - Random Forest(with GridSearchCV)

Class	Precision	Recall	F1-score	Support
Low Risk	1	0.98	0.99	171
Medium Risk	0.91	1	0.95	29
High Risk	0	0	0	0
Accuracy		0.98		
Macro Avg	0.95	0.99	0.97	200
Weighted Avg	0.99	0.98	0.99	200

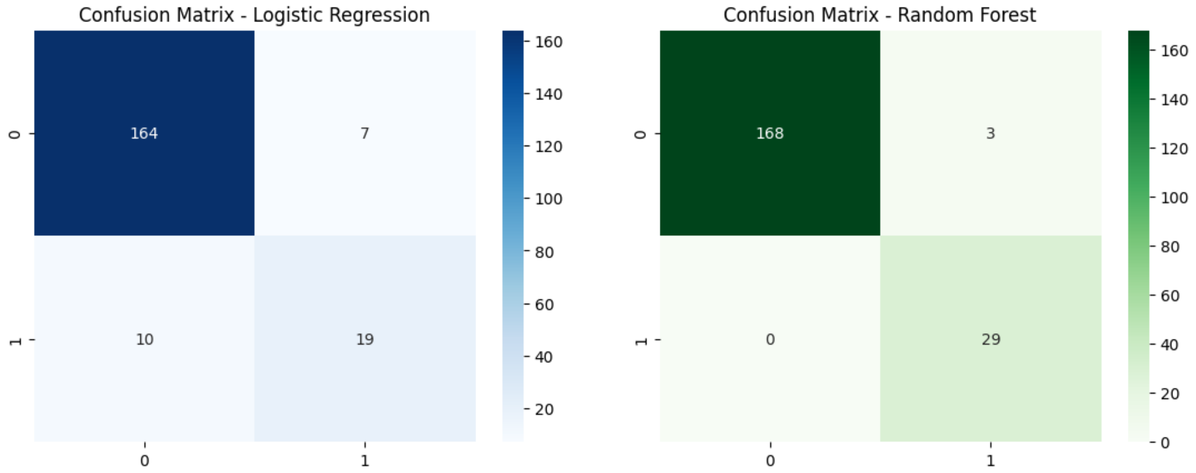


Figure 1: Confusion Matrices: Logistic Regression (Left) vs Random Forest (Right)

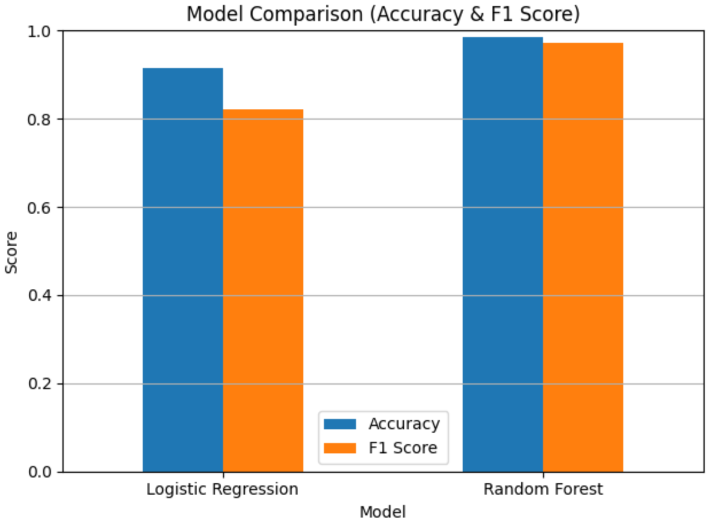


Figure 2: Model Comparison: Accuracy and F1-score

Conclusion

The comparative analysis revealed that the Random Forest model outperformed Logistic Regression in terms of accuracy, precision, recall, and F1-score across all classes. This suggests that Random Forest, with its ensemble nature, was better able to capture non-linear patterns and complex feature interactions present in the user behavior data.

The use of classification techniques for risk tiering has significant potential in real-world applications such as *credit scoring*, *fraud detection*, and *user segmentation in cyber-security*. With further enhancements—like feature engineering, tuning via `GridSearchCV`,the model's performance and interpretability can be improved.

This project not only demonstrated the viability of automated risk classification but also highlighted the importance of model selection and performance metrics in building trustworthy and actionable AI systems.