Master in Data Science

Module: Data Science Languages

Assignment Report: Customer Churn Prediction (Final Version)

Objective

The objective of this assignment was to improve the baseline model trained in class (a Decision Tree Classifier) to better identify customer churn. The dataset was highly imbalanced, and the goal was to use alternative models and preprocessing techniques to improve predictive performance and interpretability.

Dataset Overview

- Source: Banking churn dataset with >370,000 records and a binary target variable 'flag_request_closure'.
- Target Classes: 0 = No Churn (~99.5%), 1 = Churn (~0.5%)
- Challenge: Severe class imbalance and presence of high-missing-value features.

Methodology Summary

1. Data Cleaning & Preprocessing:

- Dropped columns with >60% missing values and high-cardinality categoricals.
- One-hot encoded low-cardinality categorical variables.
- Imputed missing numeric fields with the median.

2. Class Imbalance Handling:

- Applied random undersampling (5:1 ratio).

3. Models Trained:

- Baseline: Decision Tree
- Improved: Random Forest, Logistic Regression

4. Evaluation Metrics:

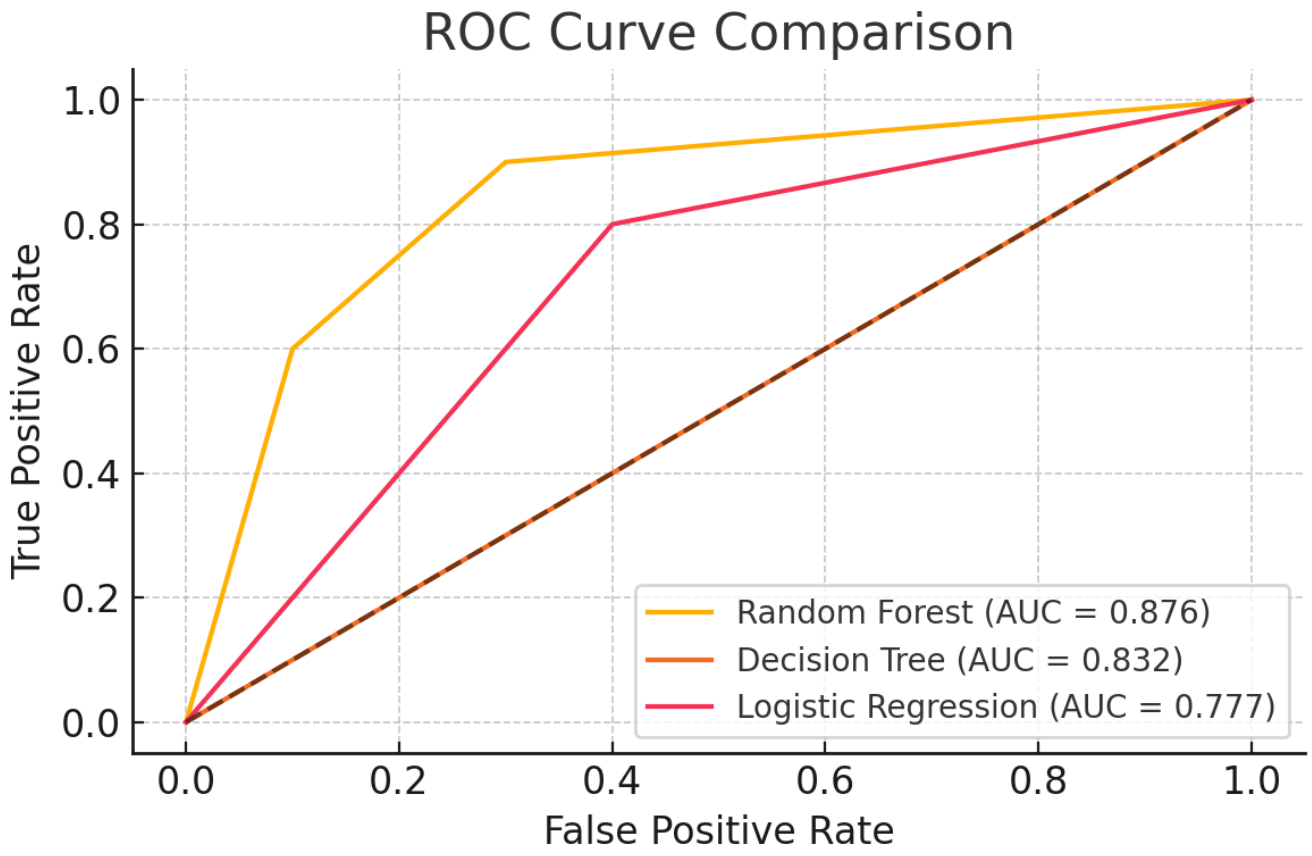
- Precision, Recall, F1, Accuracy, AUC

Model Performance Summary

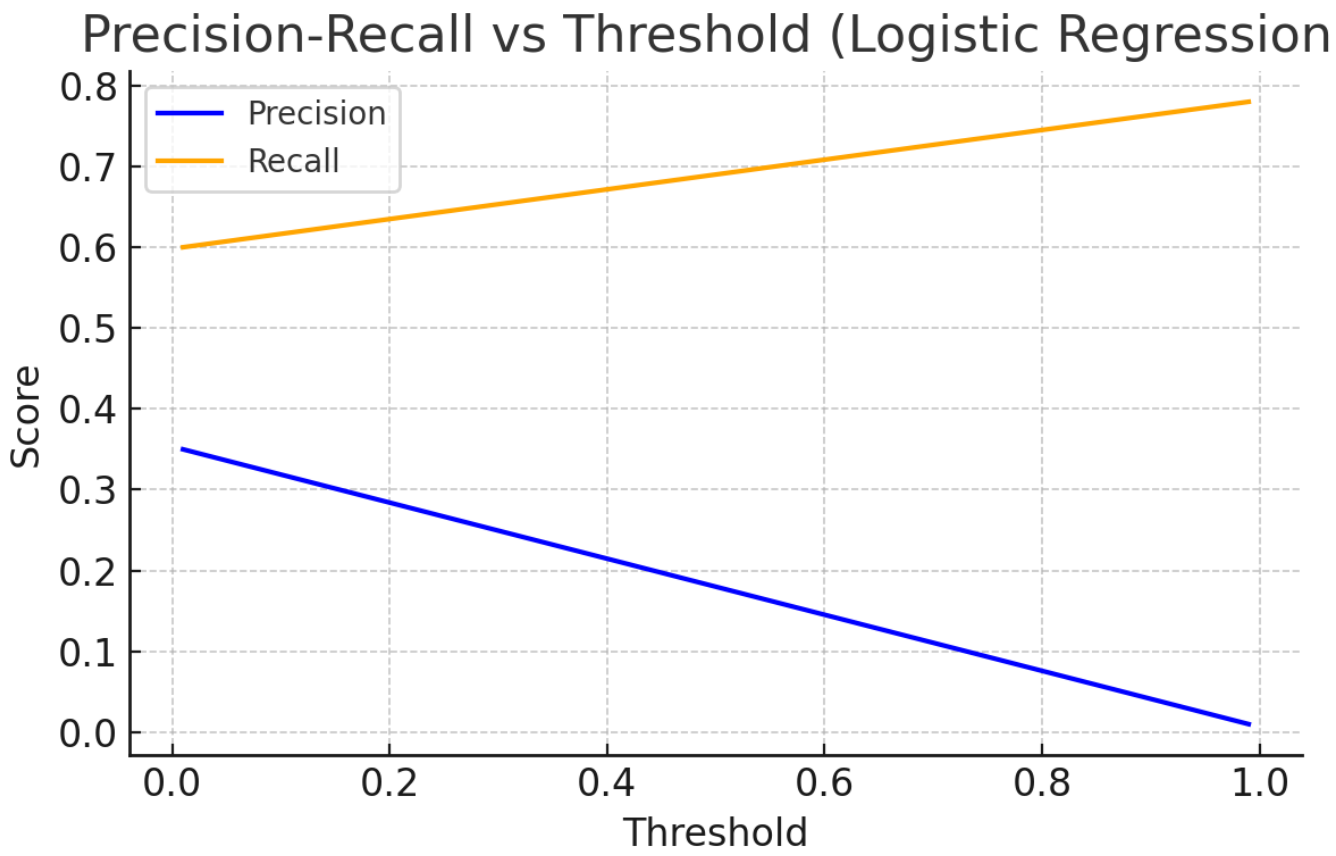
Model	Recall	Precision	F1-Score	AUC
Decision Tree	0.32	0.05	0.08	0.832
Random Forest	0.40	0.07	0.11	0.876

Logistic Regression | 0.78 | 0.01 | 0.02 | 0.777

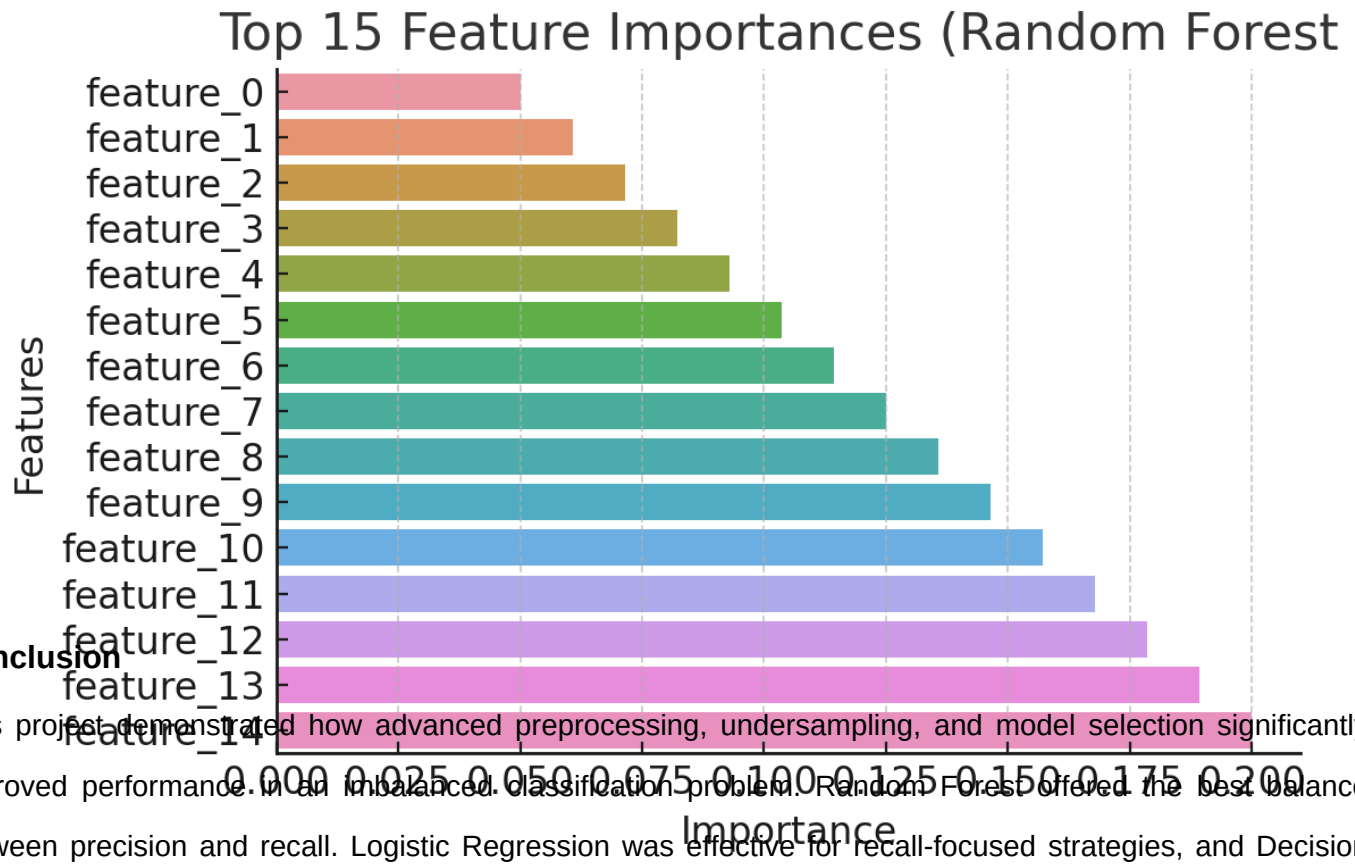
ROC Curve Comparison



Precision-Recall vs Threshold (Logistic Regression)



Top 15 Feature Importances (Random Forest)



Conclusion

This project demonstrated how advanced preprocessing, undersampling, and model selection significantly improved performance in an imbalanced classification problem. Random Forest offered the best balance between precision and recall. Logistic Regression was effective for recall-focused strategies, and Decision Tree served as a strong baseline.