Auto Loan Decision Model

Predictive Modeling for Auto Loan Applications

Mohammad Rafiqul Islam

November 20, 2024

Table of Contents

1. Introduction
2. Data Analysis
3. Modeling
4. Final Model and its Predictability
5. Conclusion

Introduction

- Enhance the application decision process for auto loans.

Introduction

- Enhance the application decision process for auto loans.
- Given Data Overview:
  - Training data around 21,000 records, test data around 5400 records
  - Both set contains 43 columns including the target variable 'bad_flag'

Introduction

- Enhance the application decision process for auto loans.
- Given Data Overview:
  - Training data around 21,000 records, test data around 5400 records
  - Both set contains 43 columns including the target variable 'bad_flag'
- Build predictive models:
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier

Introduction

- Enhance the application decision process for auto loans.
- Given Data Overview:
  - Training data around 21,000 records, test data around 5400 records
  - Both set contains 43 columns including the target variable 'bad_flag'
- Build predictive models:
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier
- Model Selection and Implementation:
  - We use classification report, ROC-AUC sroce, F1-score, and visualization
  - Implement final model to answer business questions

Data Analysis: Missing values

- Both training and test data contains significant amount of missing data

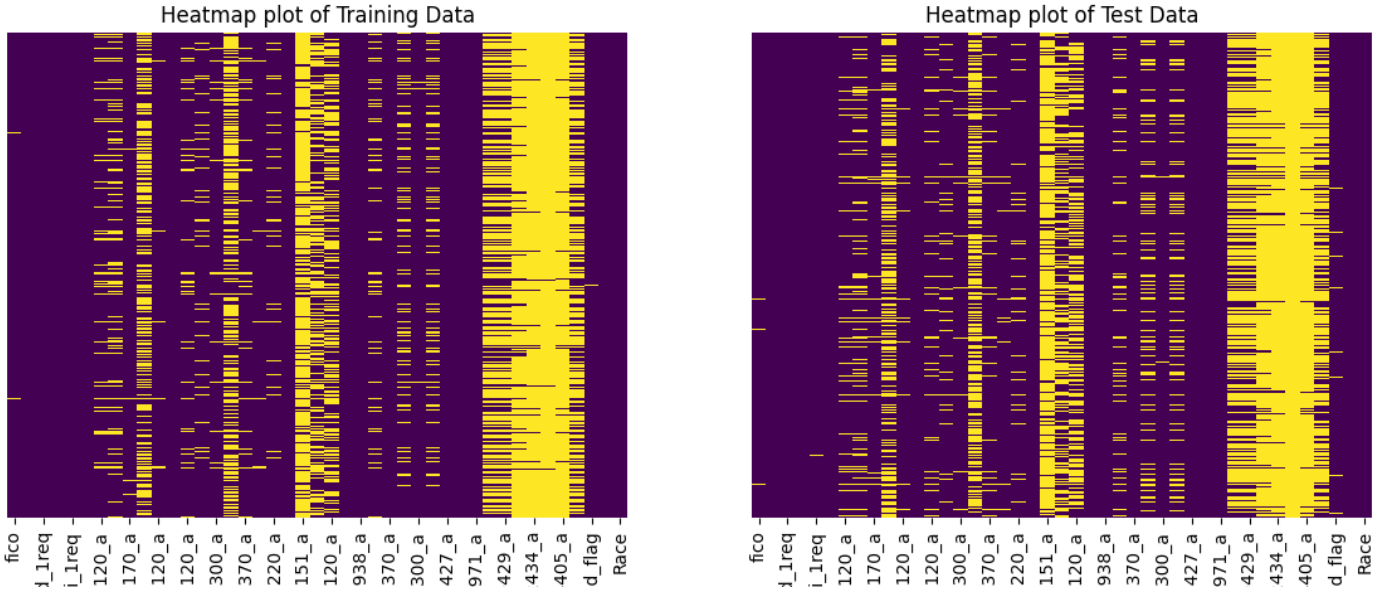

Figure 1: Missing values in train and test data

- Nine features have over 50% missing data.

Data Analysis: Missing values

- Both training and test data contains significant amount of missing data

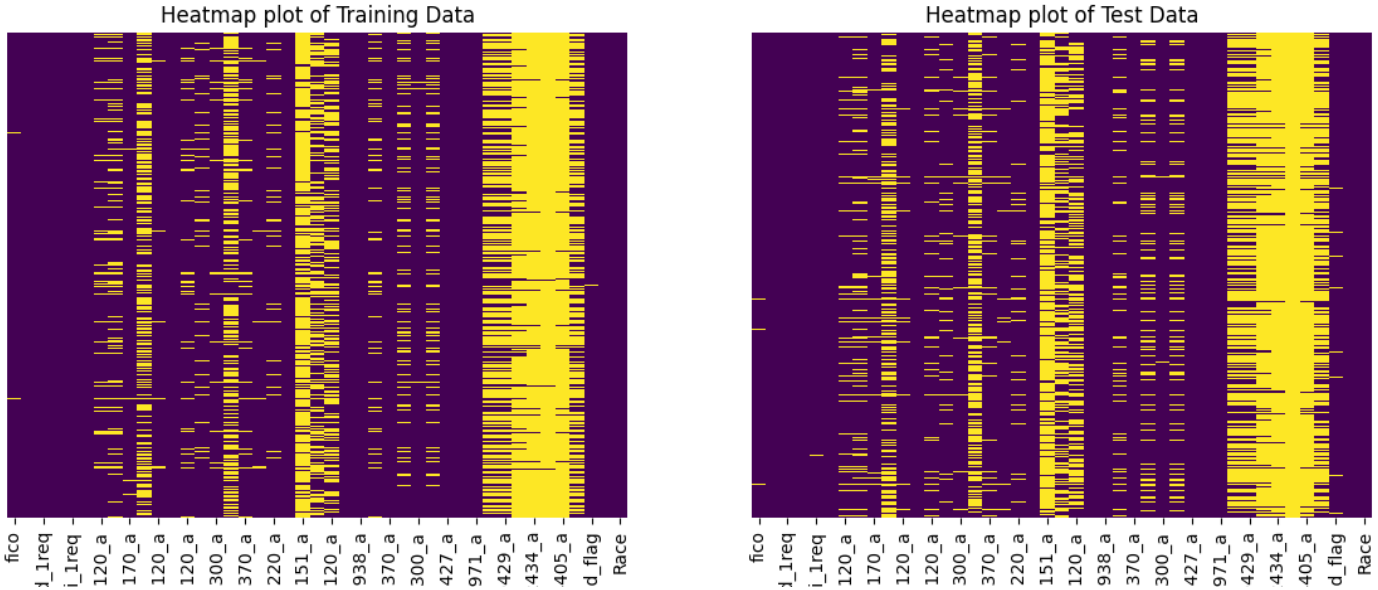

Figure 1: Missing values in train and test data

- Nine features have over 50% missing data.
- Frequency based columns: Imputed by mode

Data Analysis: Missing values

- Both training and test data contains significant amount of missing data

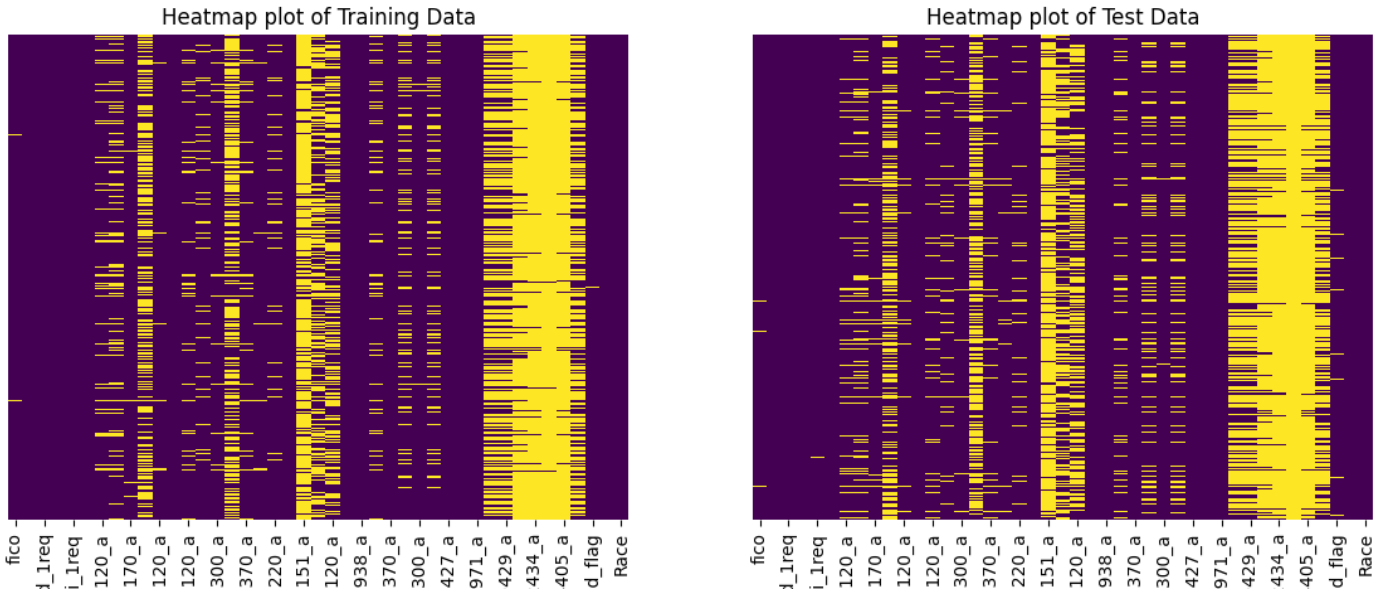

Figure 1: Missing values in train and test data

- Nine features have over 50% missing data.
- Frequency based columns: Imputed by mode
- Continuous columns: Imputed by median

Data Analysis: Exploratory Data Analysis (EDA)

- Target variable is highly imbalanced: 95.51% Poor Credit, 4.49% Good Credit

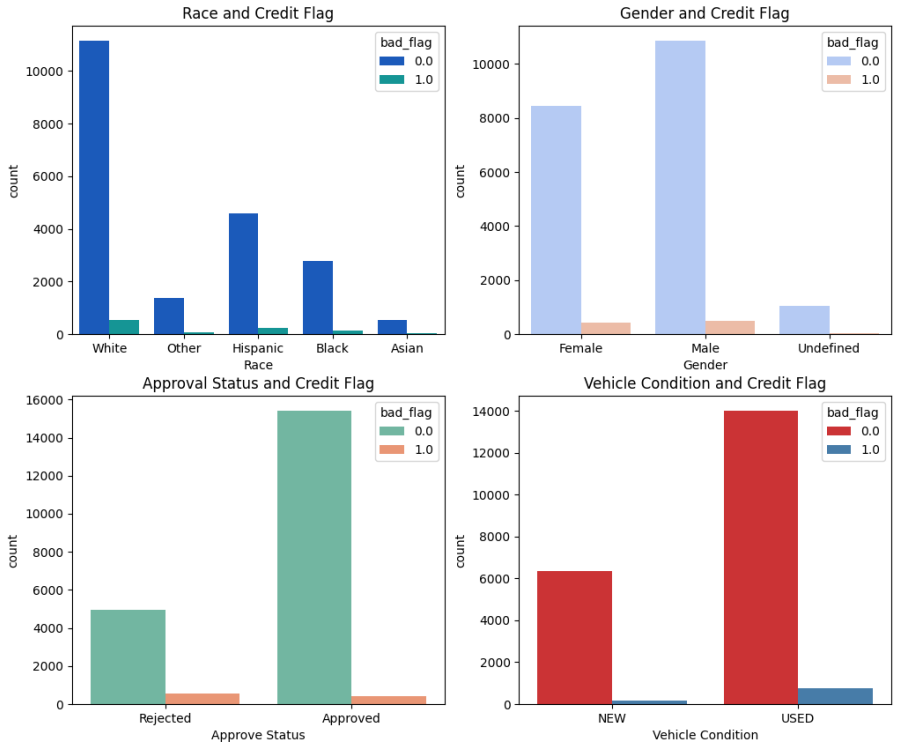

Figure 2: Bi-variate analysis of target and categorical features

Exploratory Data Analysis (cont.)

- Five frequency-based features

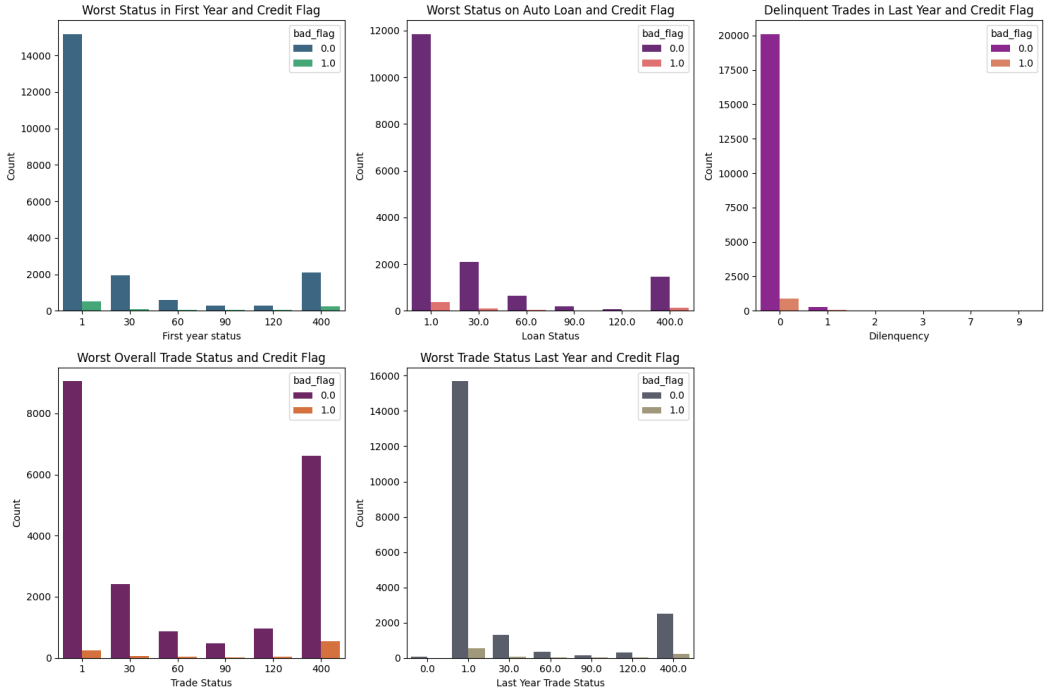

Figure 3: Bi-variate analysis of the frequency based features

- Continuous features such as BTC ratios, credit utilization rates, FICO scores were found highly impactful.

Model Overview

- Models Evaluated:
  - Logistic Regression: Linear, interpretable.
  - Decision Tree: Rule-based, prone to overfitting.
  - Random Forest: Robust ensemble method.
- Evaluation Metrics:
  - ROC-AUC
  - F1-Score
  - Classification Report

Model Performance (Without Resampling)

- Random Forest achieved the best performance:
  - ROC-AUC: 0.8078 (Mean), 0.0093 (Std).
- Classification Results (Test Data):

Metric	Class 0.0	Class 1.0
Precision	0.961553	0.268750
Recall	0.977032	0.177686
F1-Score	0.969231	0.213930
Accuracy	0.94078	

Table 1: Classification Report (Test Data)

Model Performance (With Resampling)

- Applied SMOTE for class imbalance.
- Results showed overfitting:
  - High accuracy on training data.
  - Poor generalization to test data.

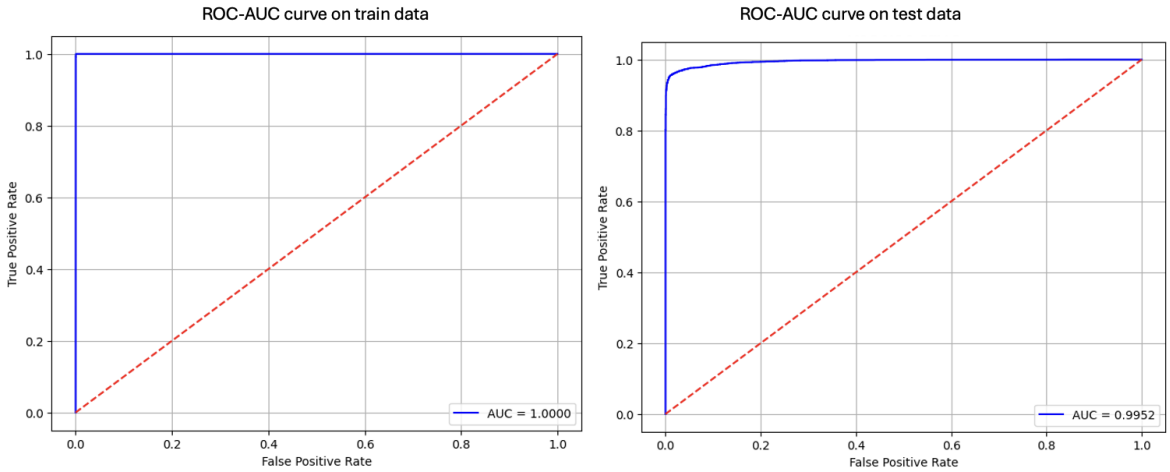

Figure 4: ROC-AUC Curve (With Resampling)

Final Model Selection

- Final Model: Random Forest Classifier.
- Key Insights:
  - Strong predictive performance.
  - ROC-AUC on unseen data: 0.94.
- Explainability:
  - Used LIME for individual prediction explanations.

Gender Equality Analysis

- Approval Rates:
  - Female: 2.20%.
  - Male: 2.68%.
  - Undefined: 5.02%.

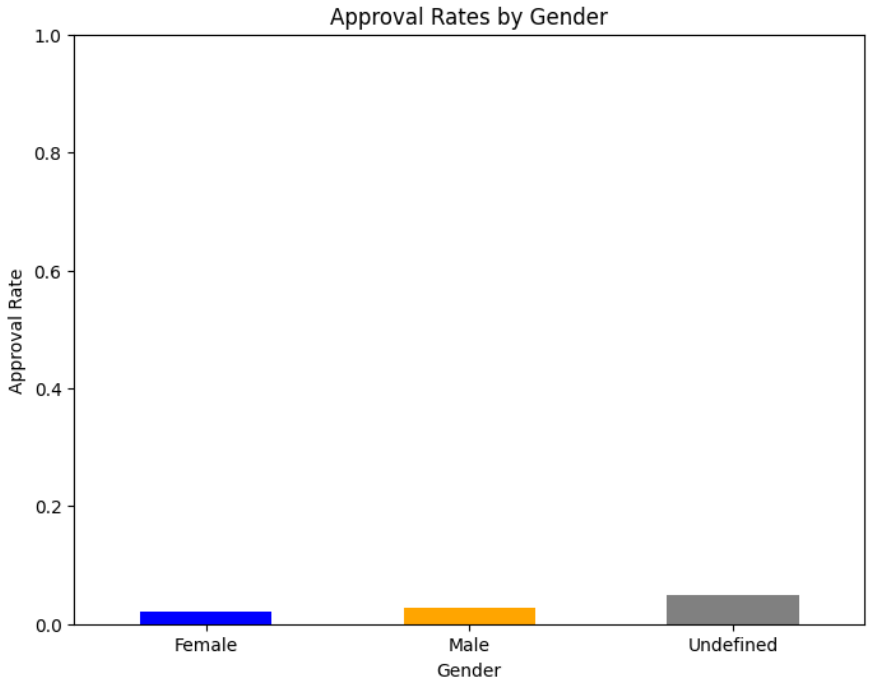

Figure 5: Approval Rates by Gender

Racial Equality Analysis

- Approval Rates:
  - Black: 2.33%.
  - Hispanic: 2.69%.
  - White: 2.59%.

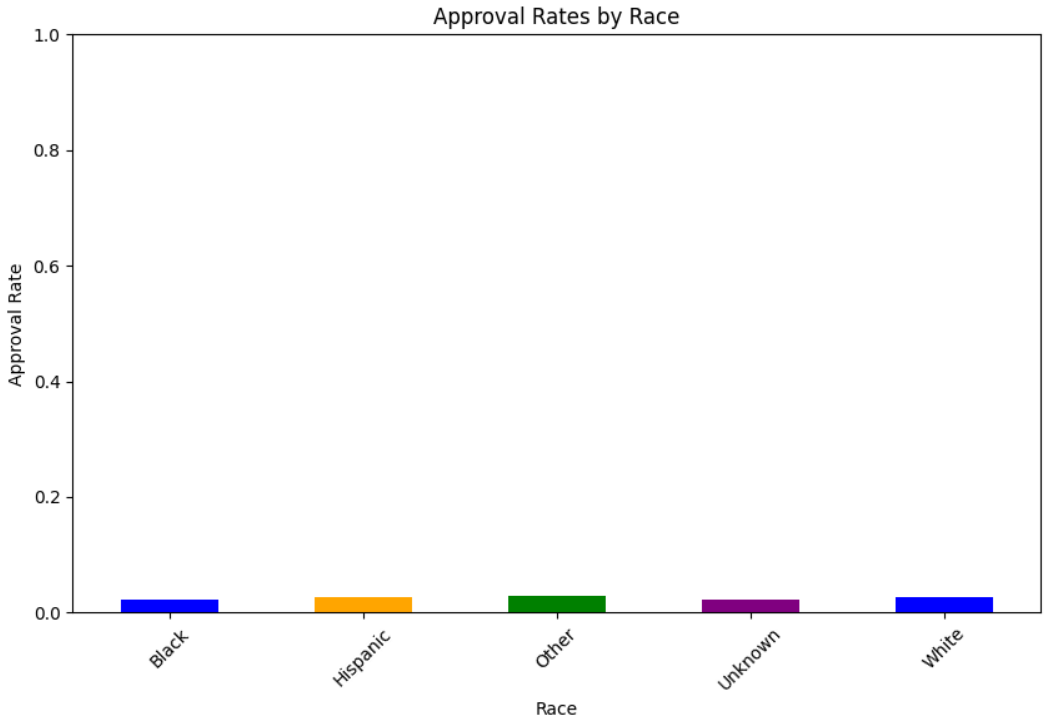

Figure 6: Approval Rates by Race

Explain a Decision

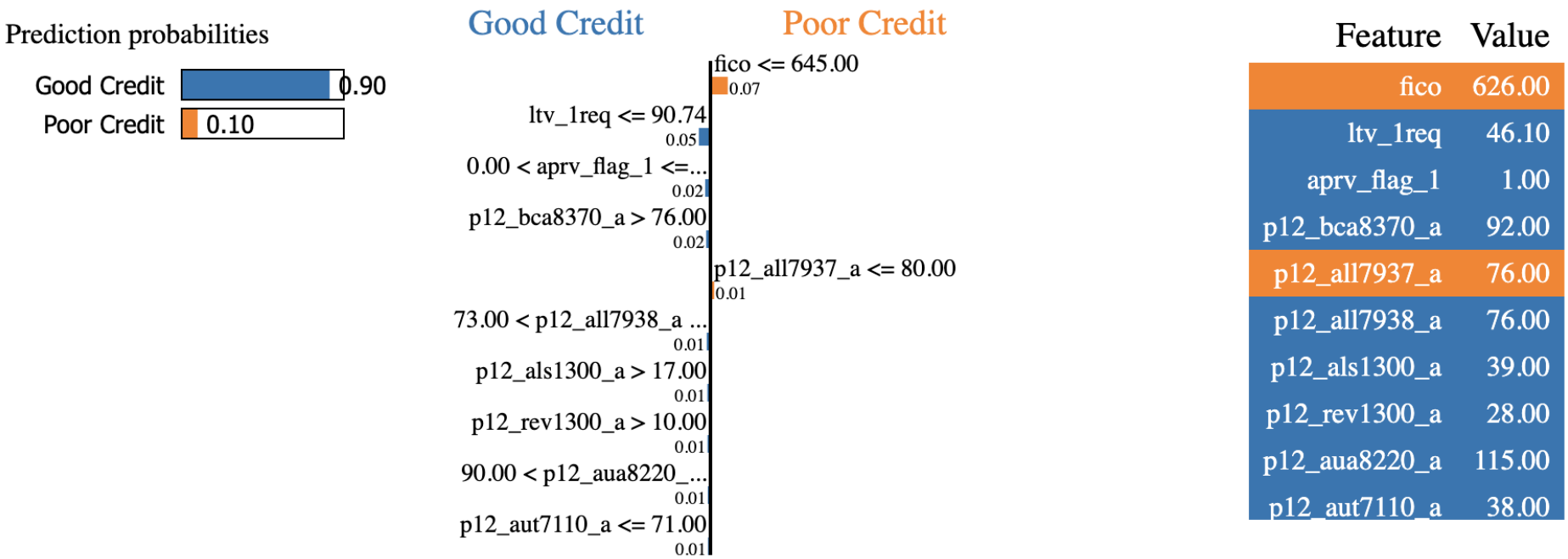

Figure 7: LIME Explanation for an Individual Prediction

Conclusion

- Random Forest demonstrated high accuracy and interpretability.
- Challenges:
  - Class imbalance.
  - Overfitting with resampling.
- Future Work:
  - Advanced models (e.g., XGBoost, SHAP for explainability).
  - Improved data cleaning and feature engineering.

Questions

Thank you!
Questions?