AI Boot Camp Project 2

Credit Risk Analysis

Team Members:
Anand Bhagwat
Usha Hariharan
Asif Mahmud
Venkat Varun Thamma

Project Overview

Project Purpose / Description

- This project showcases the application of machine learning models to analyze and predict borrower credit grades using financial data.
- Leveraging advanced datasets and methodologies provides valuable insights to support smarter and more informed lending decisions.
- Empowers institutions and borrowers

Project Overview

Goals/Problem to be solved

- Analyze/ Predict borrower creditworthiness using AI models.
- Identify the most impactful features for assessing credit risk.
- Compare the performance of multiple models to find the most accurate and efficient.

Project Overview

Data Extraction

- Kaggle ”Credit Risk Dataset”
https://www.kaggle.com/datasets/ranadeep/credit-risk-dataset/data
- Original dataset 887,000 entries rows of raw financial data
- Random Sample Size 15% of the above set—133,107 entries with 74 columns
- Key Attributes:
  - •Loan Amount (`loan_amnt`)
  - •Interest Rate (`int_rate`)
  - •Debt-to-Income Ratio (`dti`)
  - •Home Ownership (`home_ownership`)
  - •Loan Status (`loan_status`)
  - •Annual Income (`annual_inc`)

Project Overview

Data Cleaning and Transformation

- Identified columns with null values
- Reduced the dataset to only necessary columns, from 74 to 13 columns
- Created new features: `funded_amnt_to_annual_inc` and `revol_bal_to_annual_inc`.
- Changed home ownership, grade, loan status, and term to numerical values
- Defined grade and loan status as ‘y’ columns

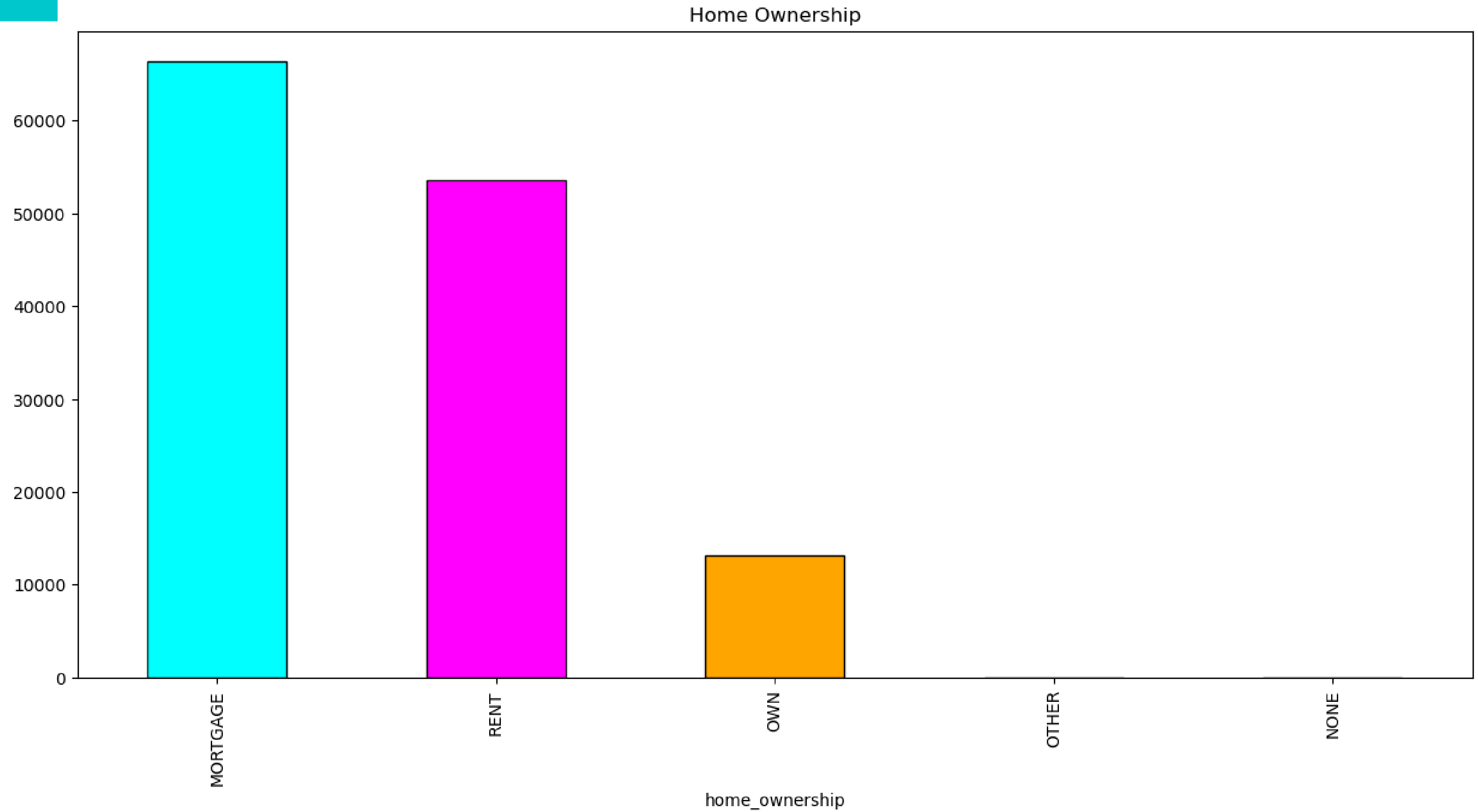
Project Overview

Overview of Exploratory Data Analysis (EDA)

- Visualized key feature distributions and their relationships to credit grades.
- Examined correlations between variables.
- Visuals on next slides:

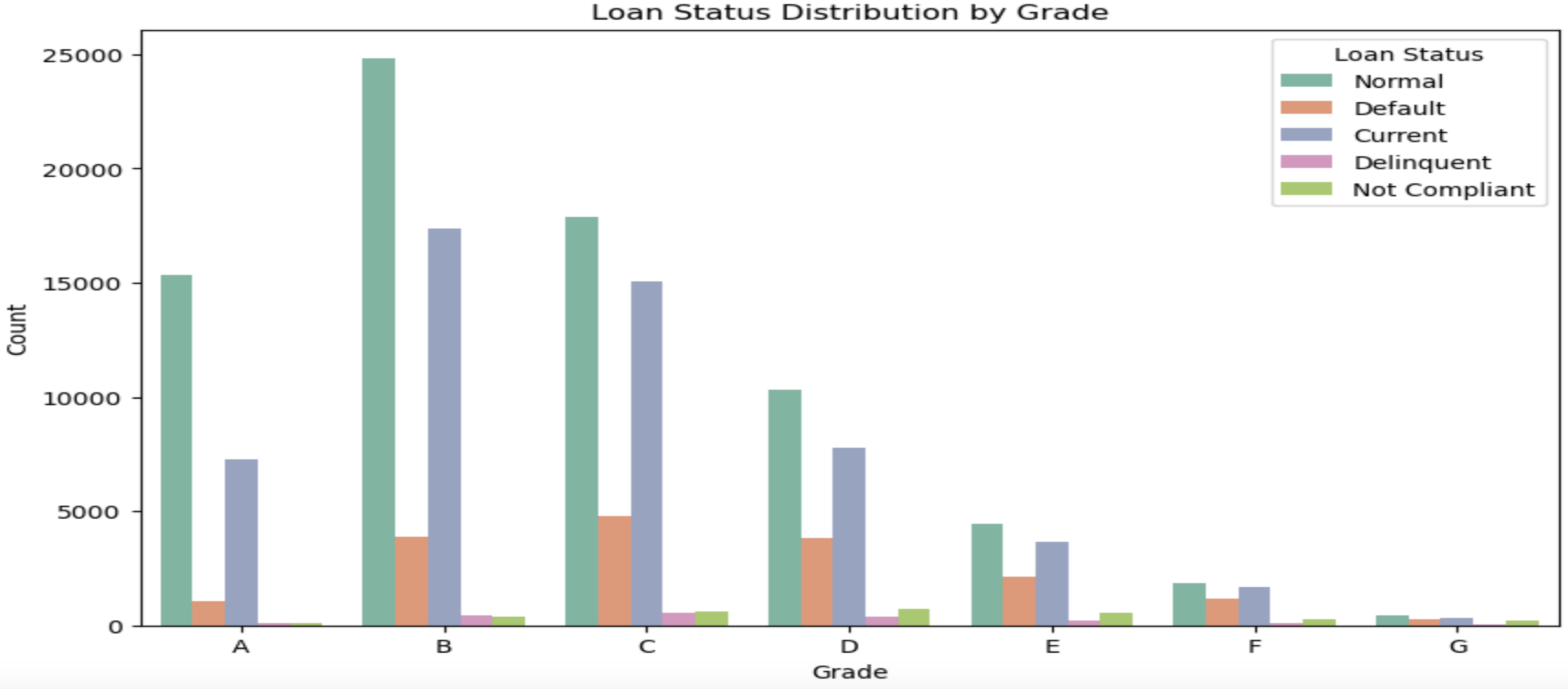
Project Overview

Dataset Count – Homeownership Type



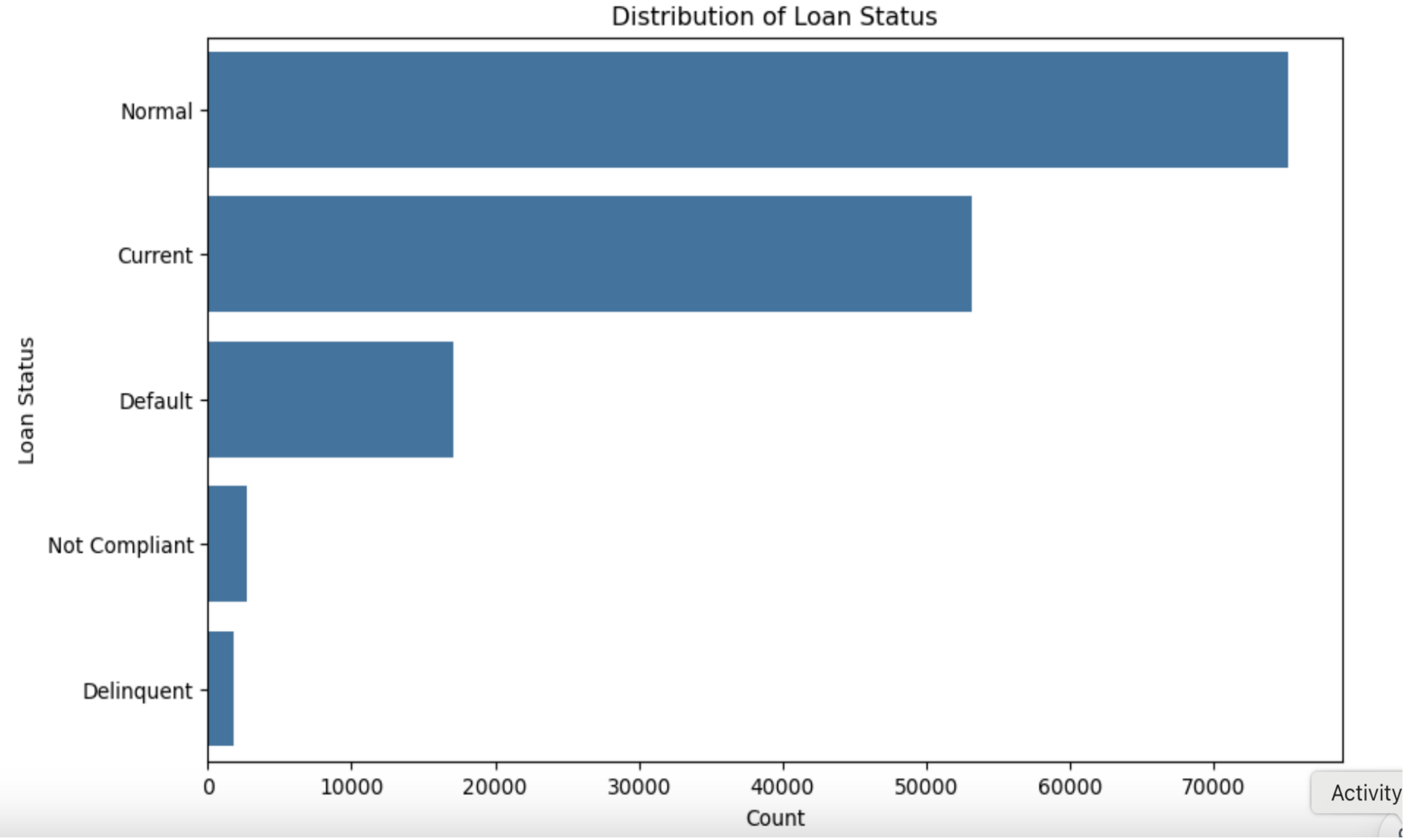
Project Overview

Dataset Count - Loan Status by Grade



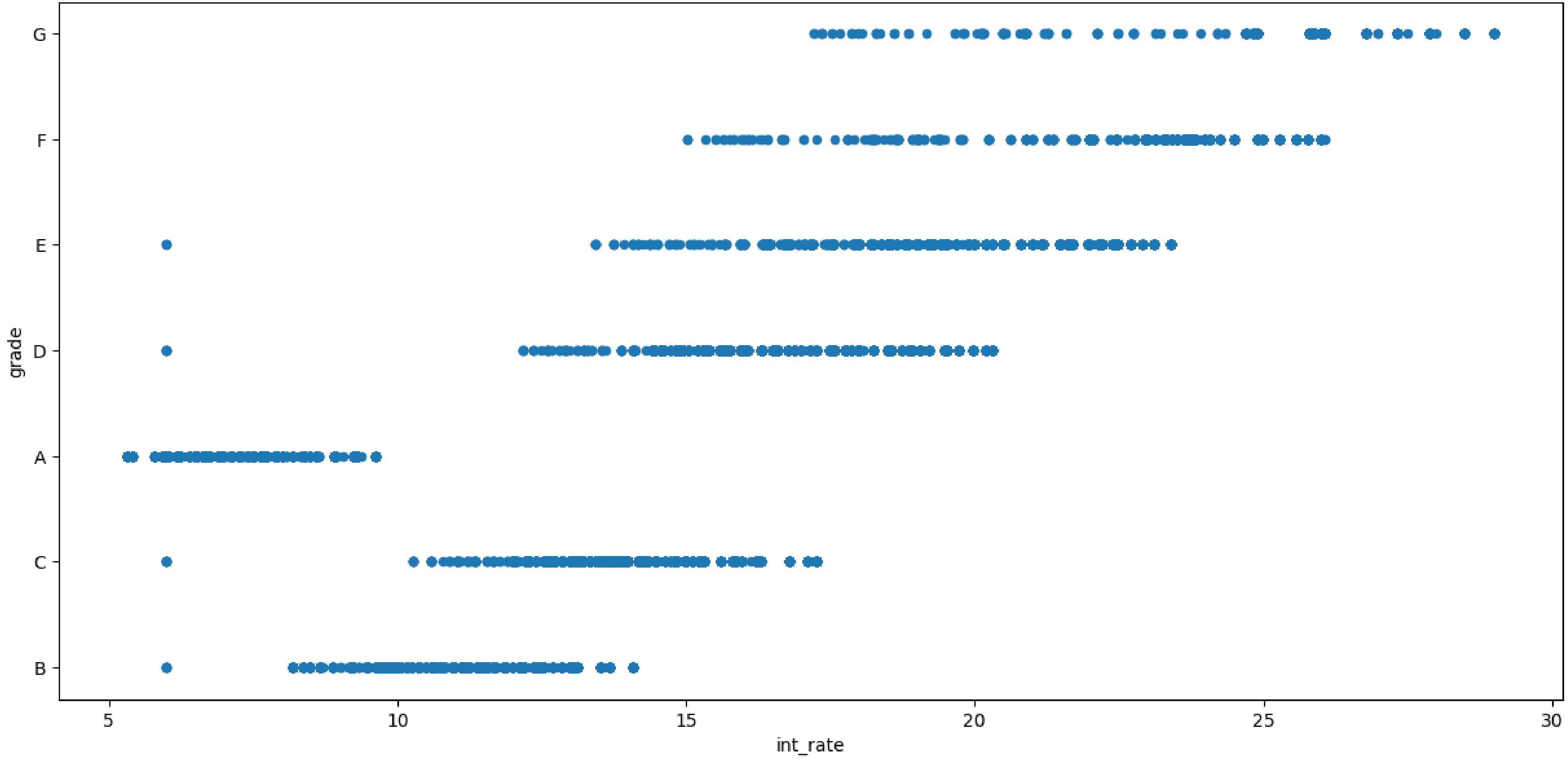
Project Overview

Dataset Count - Loan Status



Project Overview

Dataset - Grade Assignment as a function of

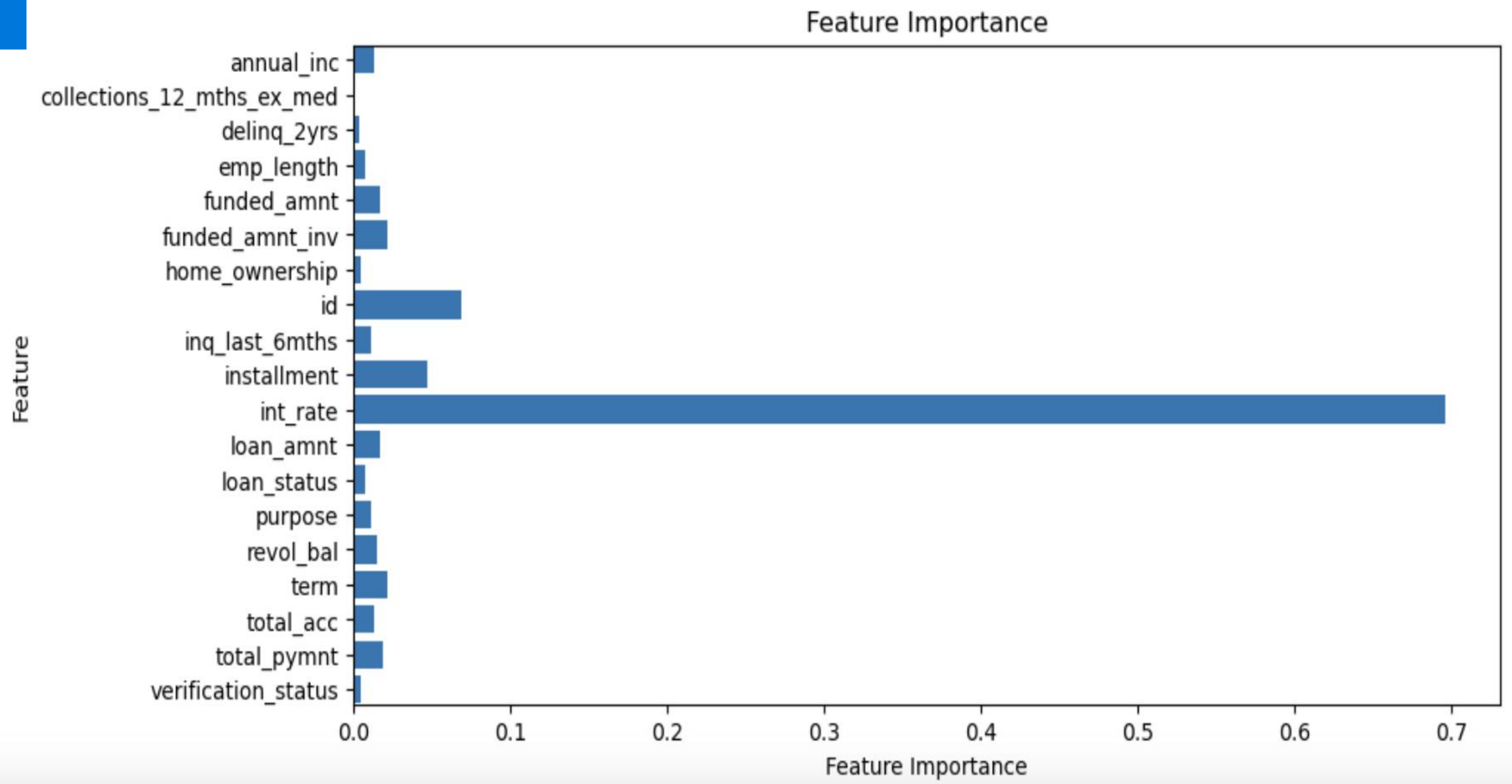


Project Overview

Approach taken to achieve goals

- Feature Importance:
- Identified top attributes such as `int_rate`, `installment`, and `loan amnt’.
- Visualized feature importance using horizontal bar plots (next slide)

Project Overview



Project Overview

Model Optimization and Evaluation

- Split data into training and testing sets.
- Scaled features using ‘Label Encoder’---ordinal variables

Project Overview

Model Optimization and Evaluation

- Trained multiple models

  1. Against y = loan_status
  2. Against y = grade
  3. Scoring ROC_AUC_SCORE

- Decision Tree: 67%

- Random Forest: 67%

- XGBoost : 89.8%

-1

Result/Conclusion 1 -- LOAN_STATUS

- MODEL PERFORMANCE: RANDOM FOREST
- Random Forest achieved a testing accuracy of 67%
- High accuracy effectively classifies “loan_status”, and can identify default risks
- High-impact features such as “int_rate” provides predictive power

Result/Conclusion 2 -- GRADE

- MODEL PERFORMANCE: XGBoost
- XGBoost**: Precision, Recall, and F1-scores close to 1.0 for all classes.
- XGBoost achieved near-perfect precision, recall, and F1-scores
- Overall accuracy rate of 99%
- Handled imbalanced classes by focusing on the most significant features

HYPERPARAMETERIZATION—LOAN STATUS

KEY INSIGHTS:

Model chosen is XGBoost – initial score 68.71%

Methods employed:

- RandomOverSampler - score 74.96%
- GridSearchCV – score 79.56%
- RandomizedSearchCV – score 89.89%

Summary

- → Machine learning model created using Grade as the determinant has a better predictor of credit risk
- → Machine learning model created using loan status as the determinant has less predictive accuracy of credit risk
- → The grade and loan status were chosen as Y values to determine borrower credit risk
- → Grade provided better accuracy than loan_status

Problems Encountered

1. Data Quality: Size of data sets were large and difficult for Github to process for expedient analysis

2. Class Imbalance: Resulting from XGBoost, same number of components were not used in classification report for 0 and 1 values

3. Even with n_estimators = 200, the accuracy for loan status was subpar

4. High Dimensionality: Selection of models to utilize was difficult; for example, SMOTE was dropped due to poor results, it also cannot process NaN values, this changed all the scores

Future Considerations

FUTURE ML ENHANCEMENTS:

1. Hyperparameter Tuning: Incorporating hyperparameter optimization for Random Forest and XGBoost to improve performance further.

2. Additional Datasets: Our models can be further validated with additional datasets to ensure they generalize well with other types of loans

3. Real-time Dashboard: Live dashboard can provide real-time credit scoring and loan risk assessments

Future Considerations

FUTURE FORWARD CREDIT CONSIDERATIONS:

1. Emerging / alternative data sources: social, connected device data
2. Environmental/social sustainability metrics: spending pattern alignment with ethical and community welfare practices
3. Blockchain and decentralized credit scoring: scoring metrics based on smart contracts, blockchain transactions, and wallet activity
4. Real-time and Dynamic scoring: real-time analysis of transactions and adjustments for critical life events
5. Ethical and Inclusive models: bias-resistant models, creating nontraditional scoring models that support credit extension to underserved populations
6. Network-based creditworthiness: peer influence models, group creditworthiness
7. Hybrid models: FICO+ ML derived alternative data

Future Considerations

FUTURE FORWARD CREDIT CHALLENGES:

Key Challenges to Address

While these paradigms offer tremendous potential, they also present challenges:

•**Data Privacy and Security**: Safeguarding sensitive consumer data.

•**Fairness and Bias Mitigation**: Ensuring equitable access and avoiding systemic bias.

•**Regulatory Compliance**: Aligning innovations with strict regulatory frameworks.

•**User Trust**: Building confidence in nontraditional scoring methods.

Exploring these paradigms in credit risk can revolutionize financial inclusion, improve risk accuracy, and cater to the evolving financial ecosystem.