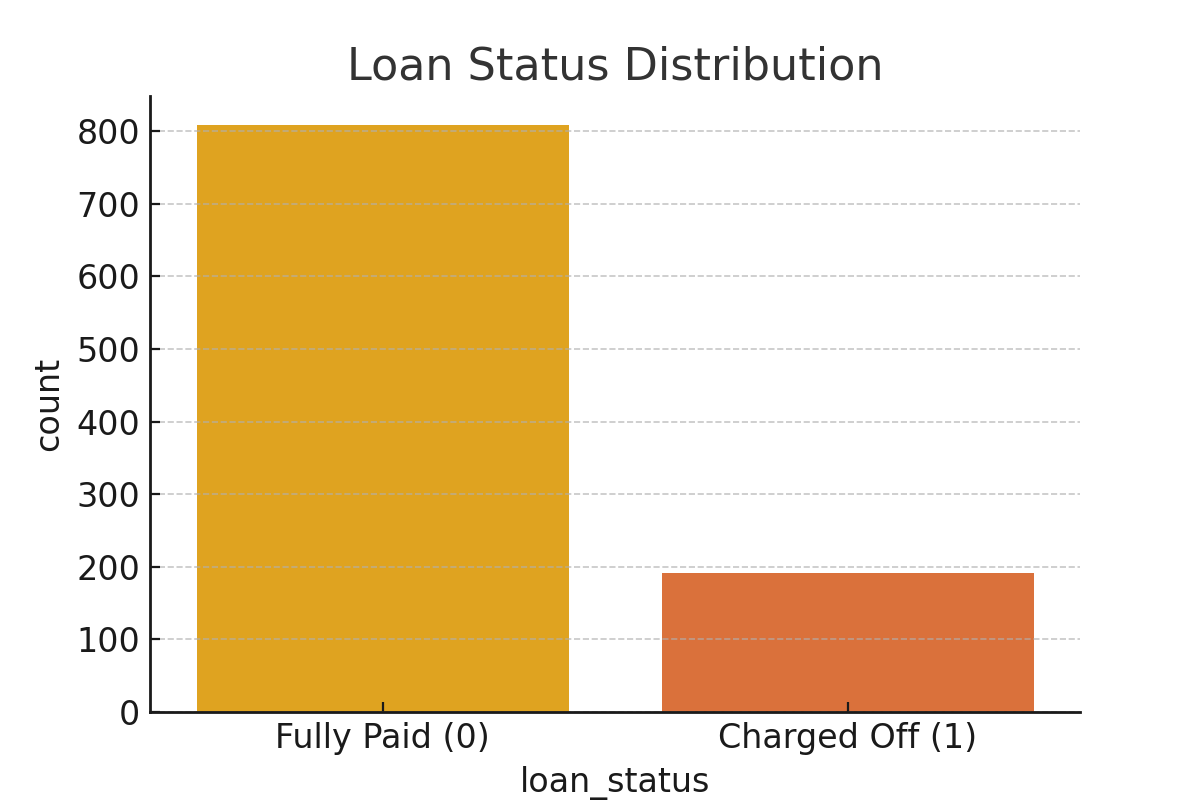
Exploratory Data Analysis (EDA) Report

# 1. Introduction

This report presents the Exploratory Data Analysis (EDA) for the Credit Risk Prediction project using the Lending Club dataset. The aim is to understand the data structure, identify important features, and highlight data characteristics such as class imbalance.

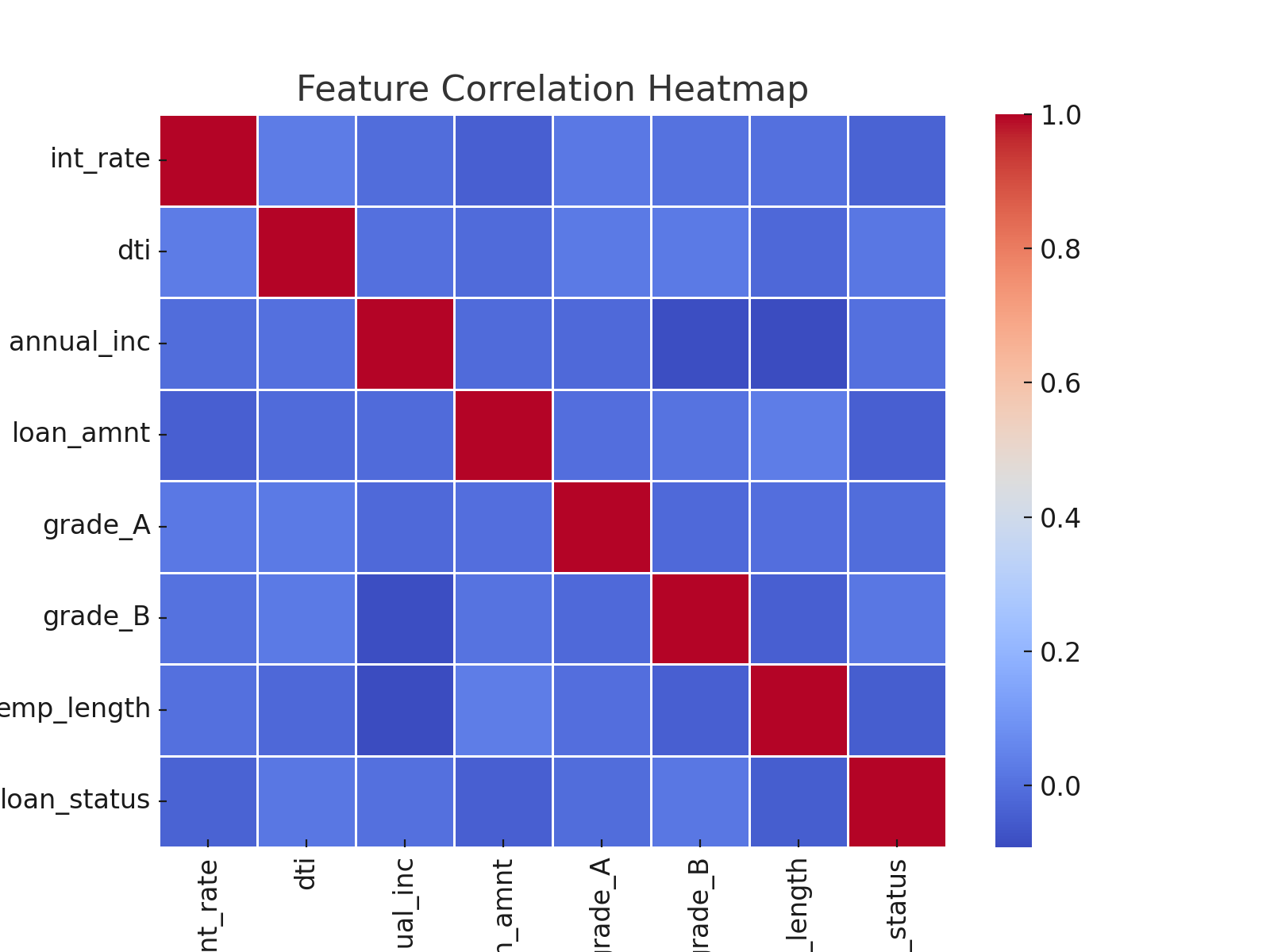
# 2. Class Distribution

The dataset exhibits a class imbalance between loans that were 'Fully Paid' and those that were 'Charged Off'. This imbalance can affect model performance, necessitating techniques like SMOTE or class weighting.



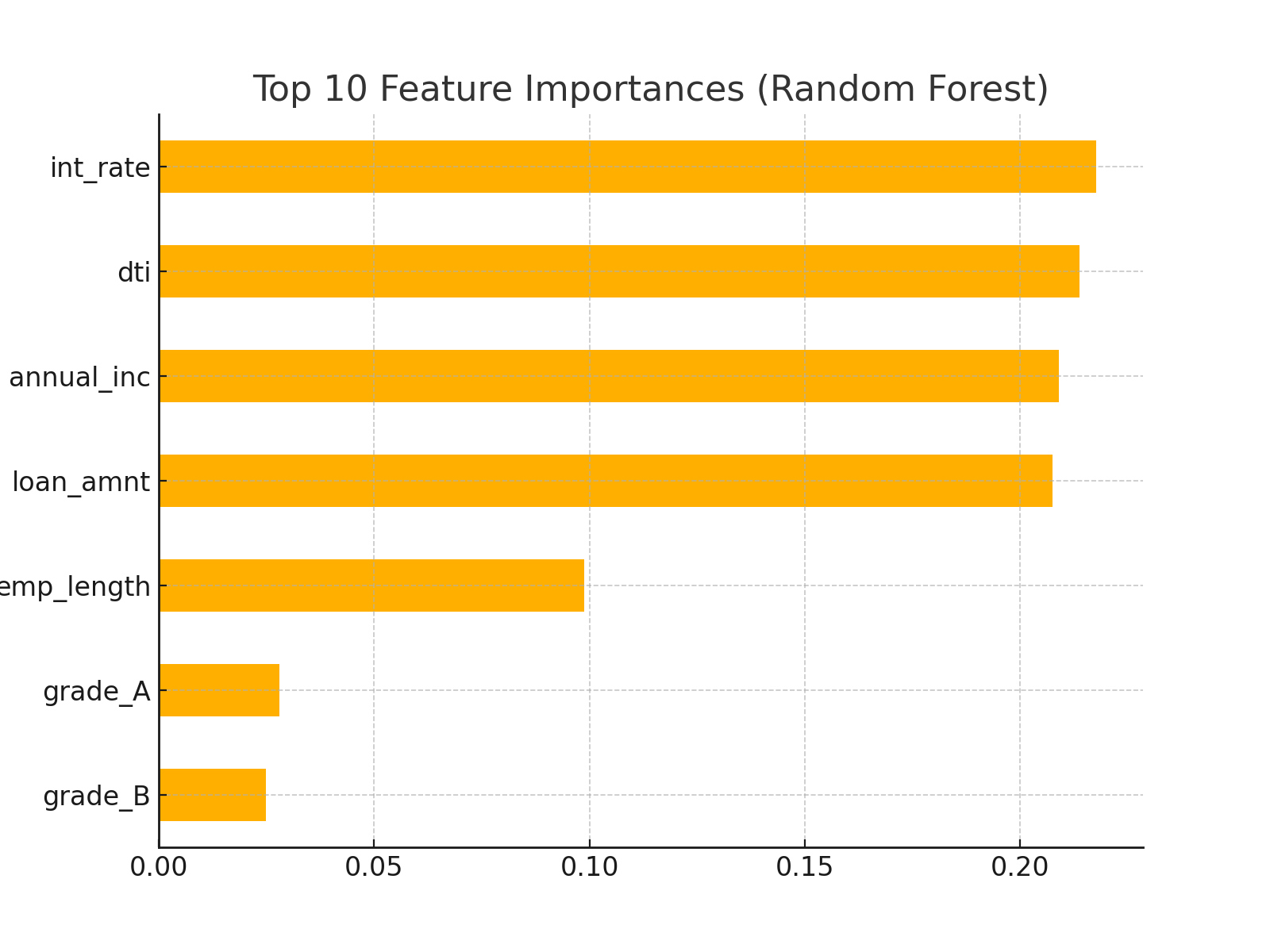
# 3. Feature Correlation Heatmap

Correlation analysis helps identify multicollinearity among numerical features. This step is crucial for avoiding redundant information and ensuring model stability.



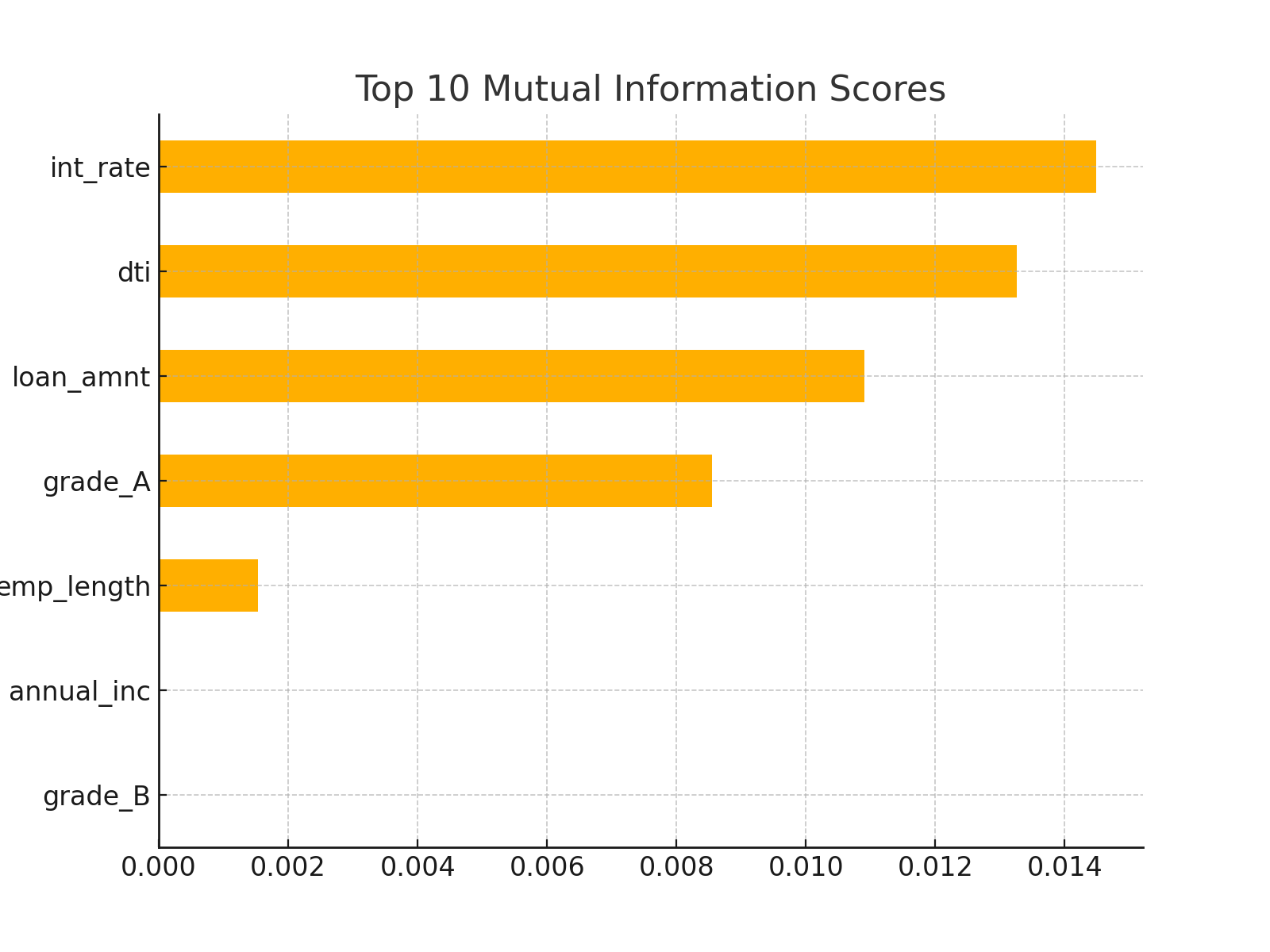
# 4. Random Forest Feature Importances

Random Forests provide insight into feature importance based on their contribution to reducing impurity. This helps prioritize features for model training.



# 5. Mutual Information Scores

Mutual Information (MI) captures both linear and non-linear dependencies between features and the target variable. MI is useful for selecting relevant features that may not be evident from correlation analysis.



# 6. Summary

The EDA reveals significant class imbalance and highlights key predictive features such as 'int\_rate', 'dti', and 'loan\_amnt'. These insights guide the feature selection and modeling process, ensuring better performance and interpretability in the credit risk prediction task.