LOAN DEFAULT ANALYSIS AND PREDICTION

Introduction

What is a default?

Default is the failure to make required interest or principal repayments on a debt, whether that debt is a loan or a security.

What is the source of data?

- Lending Club's loan data which consists of over 3M loan applications with a default rate (‘Charged Off’) of ~20% & with over 140+ predictors.
- Variables include customer's credit background, demography, loan specifications, borrower’s employment history, etc.

VARIABLE	DESCRIPTION
grade	LC assigned loan grade
addr_state	The state provided by the borrower in the loan application
installment	The monthly payment owed by the borrower if the loan originates.
annual_inc	The self-reported annual income provided by the borrower during registration.
issue_d	The month which the loan was funded.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
loan_status	Current status of the loan.
int_rate	Interest Rate on the loan.
id	A unique LC assigned ID for the loan listing.
…	…

Steps

- Data Cleaning
- EDA
  - Distribution of X variables (Predictors) - Univariate Analysis
  - Relationship between Predictors - Multivariate Analysis
  - Relationship between the Y and X variable (Response and Predictors)
- Feature Selection and Transformations
- Modeling
  - Logistic Regression
  - Random Forest
  - XGBoost
  - SVM

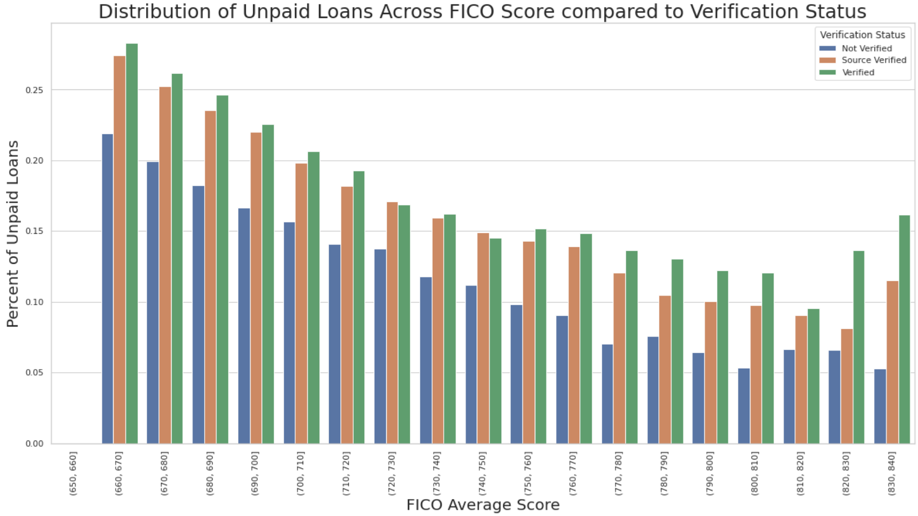
Data Cleaning

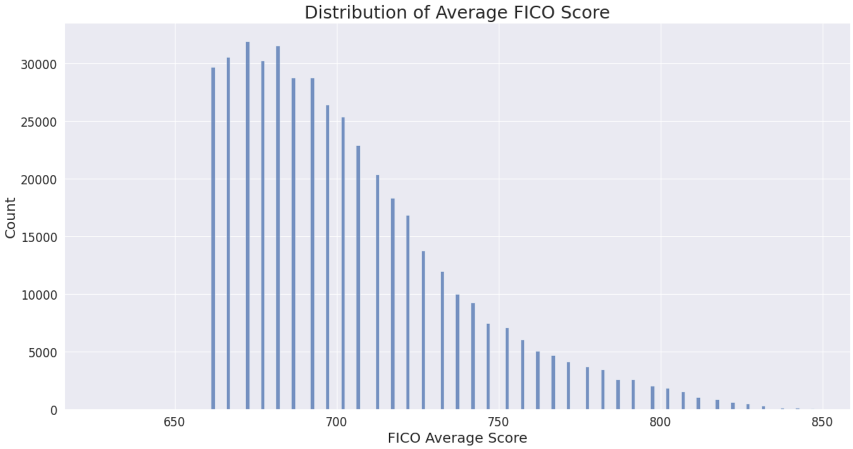
- **Cleaning Columns**
  - Removed Missing values which had more than 95% null values
  - Dropped variables,if they had more than 80% same values
  - Removed variables that had high correlation
  - Irrelevant variables were removed
  - Dropped variables that were available after the loan was sanctioned
- **Cleaning Rows**
  - Rows having duplicate entry were checked and were removed
  - Data types were converted to int, float or boolean(for xg-boost)
  - For all models, only Fully paid, Charged Off and Default Loans were left
- **Cleaning Results:**
    - Number of Variables Left: 47

Exploratory Data Analysis: Data Summary

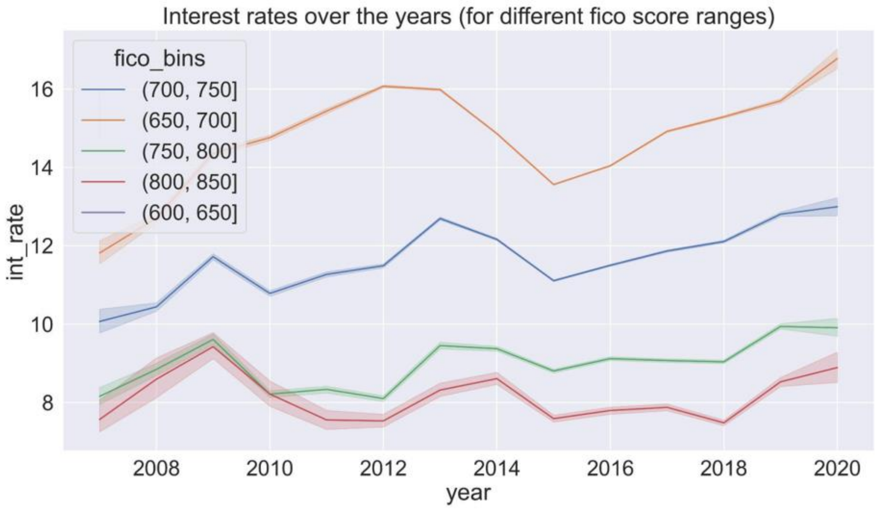
Number of Observations (Completed Loan Applications)	1860764
Number of Predictors	141
Number of Defaults	19.51%
Number of Categorical Columns	35
Number of Numeric Columns	106

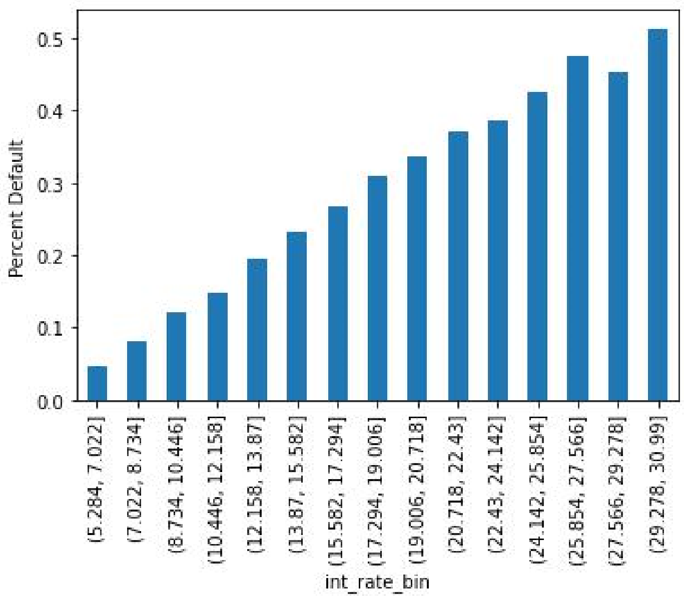
EDA:Predictor Analysis, FICO Score





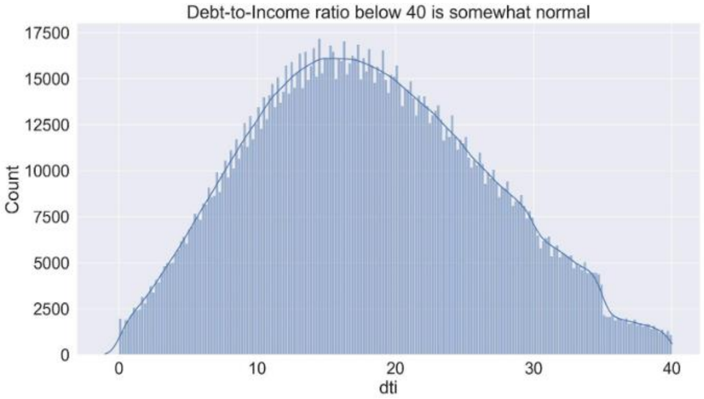
EDA: Predictor Analysis, Interest Rate

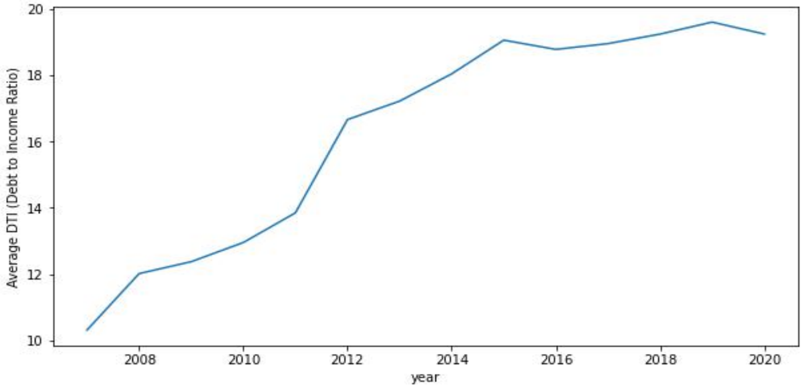




Exploratory Data Analysis:DTI

DTI-Debt to Income Ratio

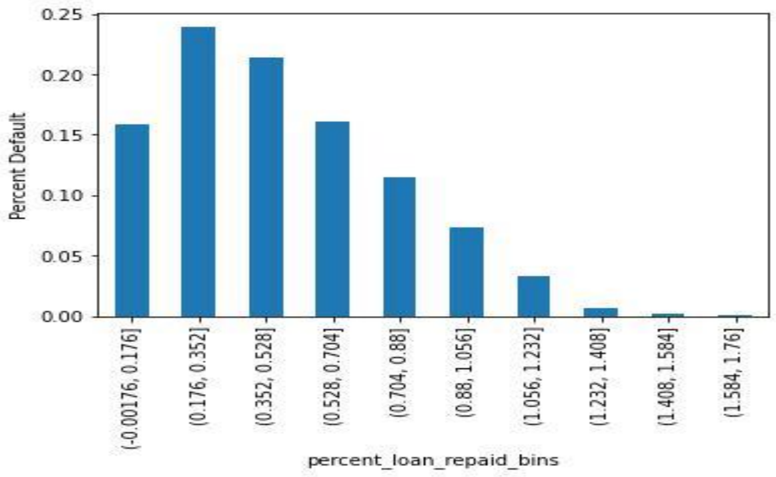


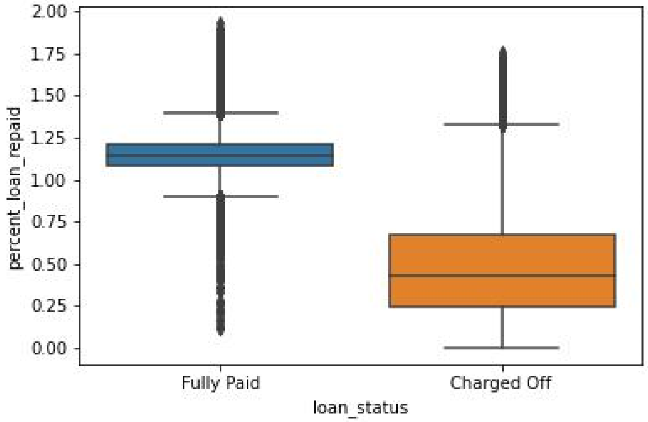


Average DTI (Debt to Income) ratio of the applications at Lending Club is increasing over the years, i.e., Lending Club is accepting riskier loan applications over the years

Exploratory Data Analysis: **Percentage Repaid**

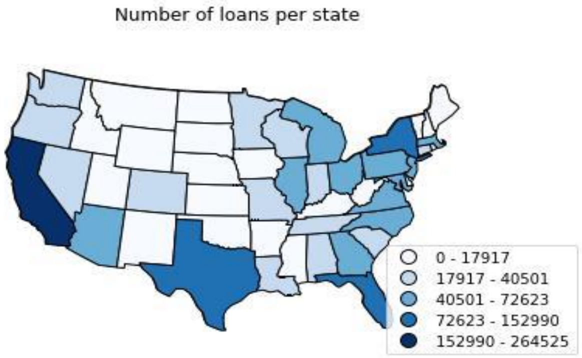
Relation between percentage repaid with Loan status

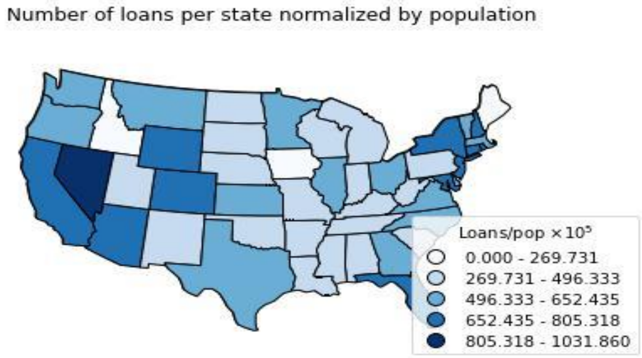




Higher the percentage of loan repaid, lower the chances of defaults.

Exploratory Data Analysis:State-Wise

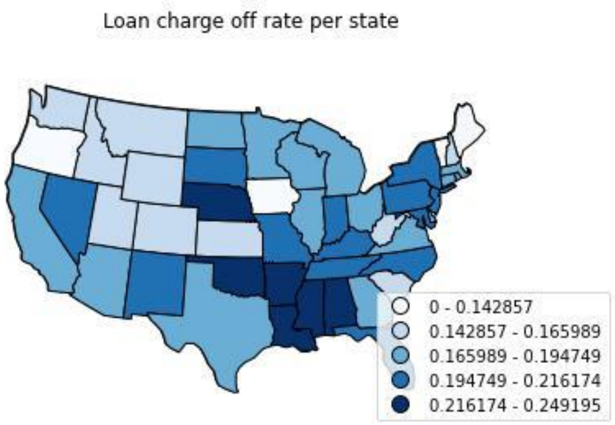


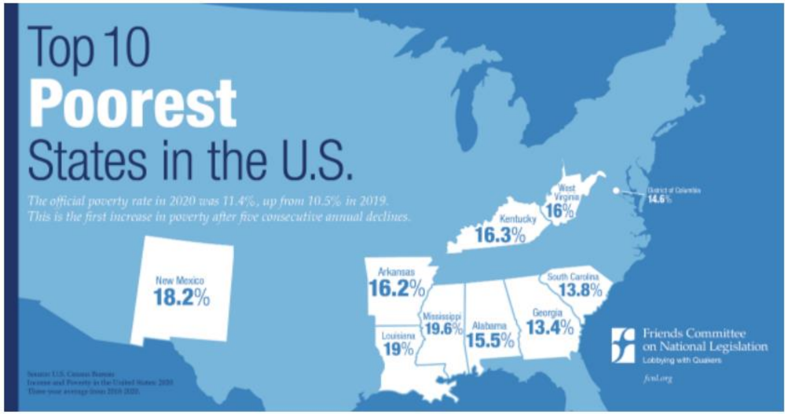


California had the maximum amount of loan sanctioned.

Nevada had the most amount of loan sanctioned when normalized with population.

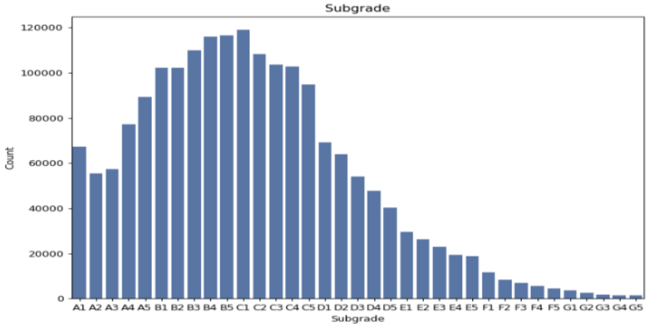
Exploratory Data Analysis:State-Wise

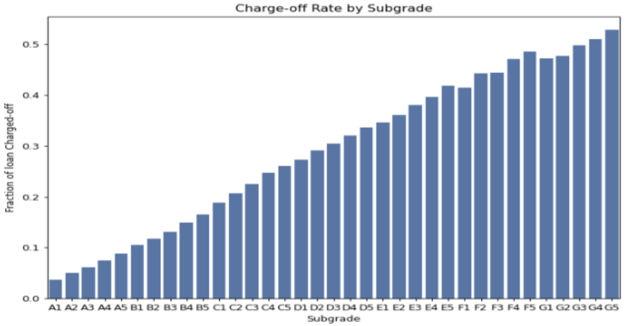




South-eastern states like Louisiana,Alabama and Mississippi had the highest amount of loan defaults.

Exploratory Data Analysis:Grading





The charge-off ratio shows a uniform trend which increases from A1 to G5.It reflects that the applicant grading by the lending club is quite accurate.

Modelling Approach

Objective: Compute important features for predicting a Loan Default

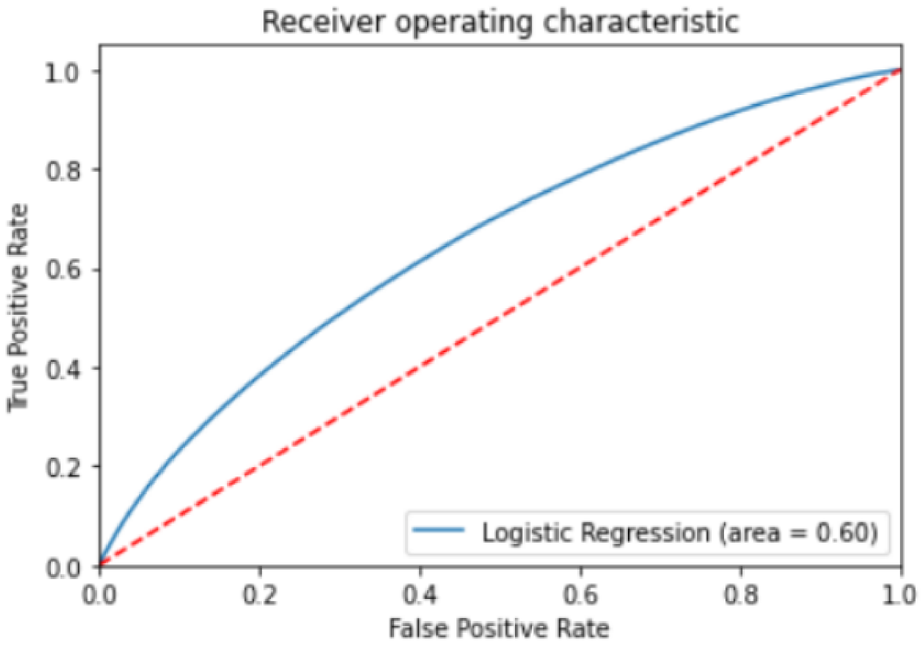
Approach:

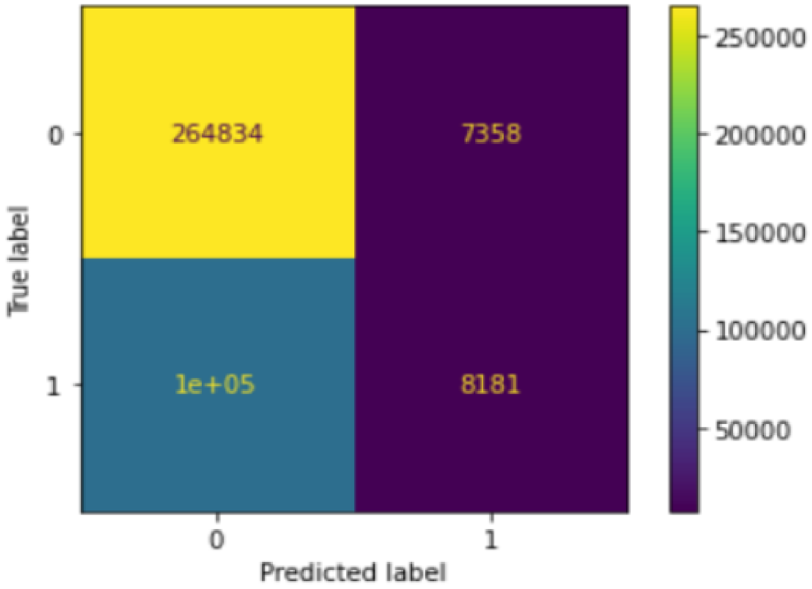
1. Select the statistically significant variables (p-value < 0.05) using logistic regression
2. Run Random Forest, XGBoost & SVM classifiers to compute the feature importances

Evaluation Metric: In this use case, classifying loan default is as important as classifying the non-loan defaults. Hence, we used **AUC ROC** metric as our evaluation metric

Handling Class imbalance: Loan defaults comprises of 20% of the total approved applications. Since we have significant number of records, we used random undersampling to handle the class imbalance problem by equally distributing the 0s & 1s (i.e. 50%/50%)

Modelling Results - Logistic Regression

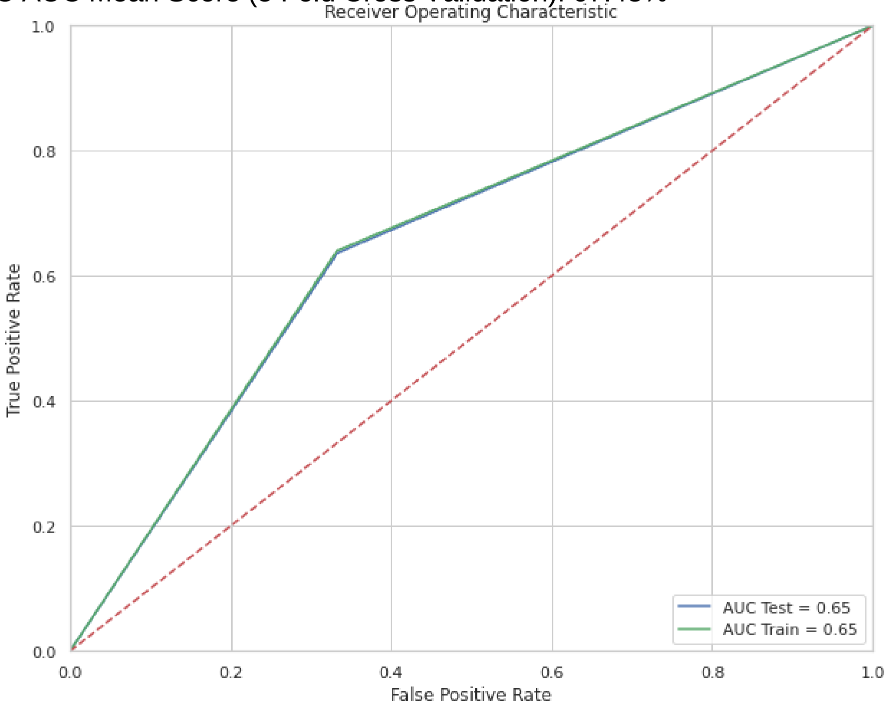


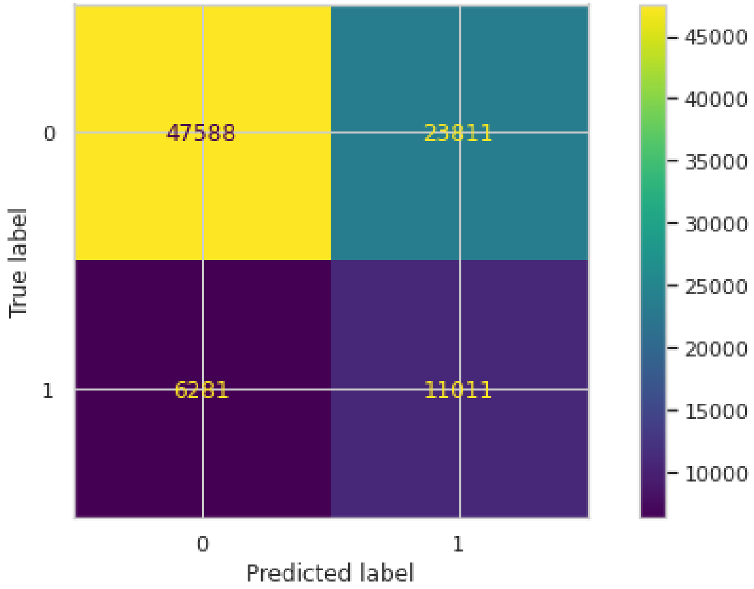


Using 95% confidence interval, we dropped ~20% of the variables using LR

Modelling Results - Support Vector Machines

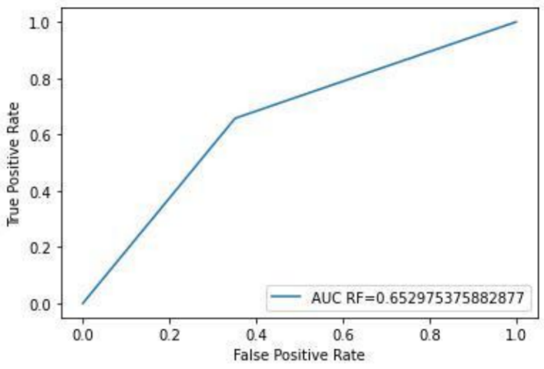
ROC AUC Mean Score (5 Fold Cross Validation): 67.43%

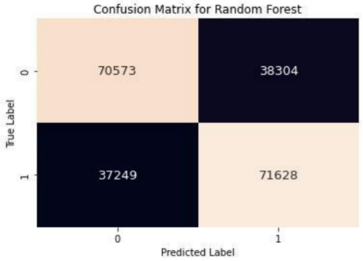


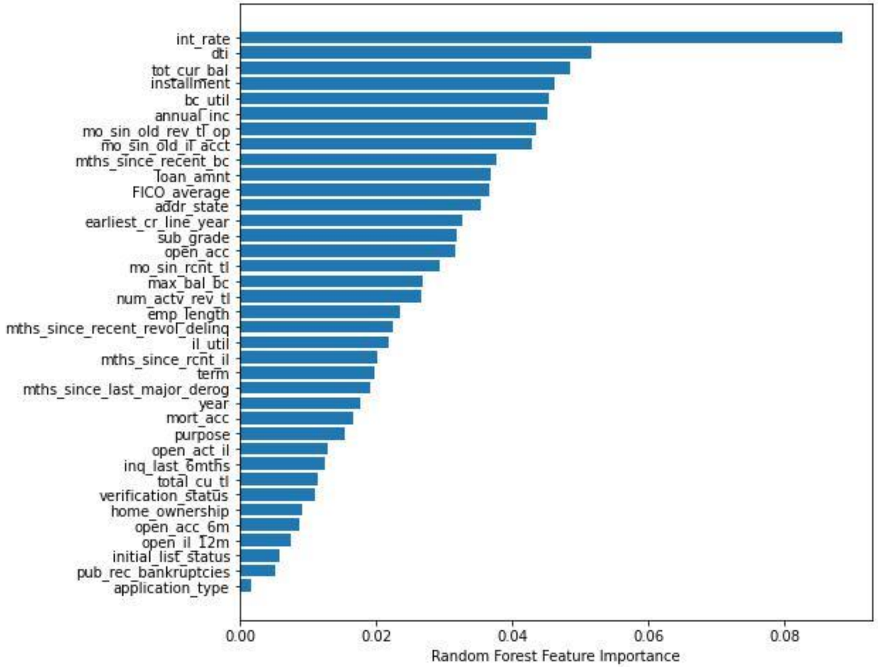


Modelling Results - Random Forest

2. Random Forest: ROC AUC Score: 65.29%

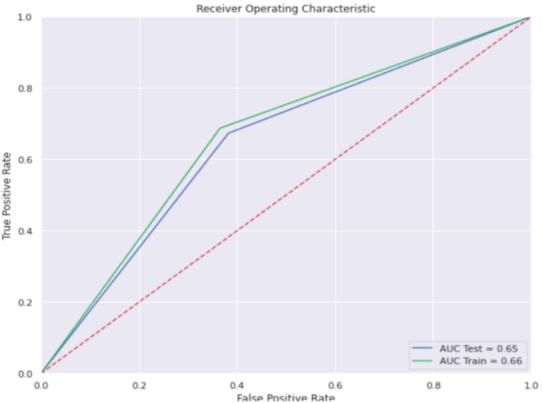


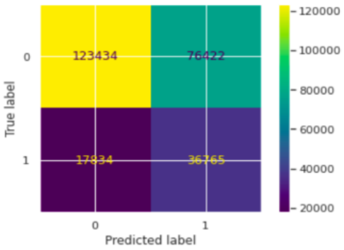


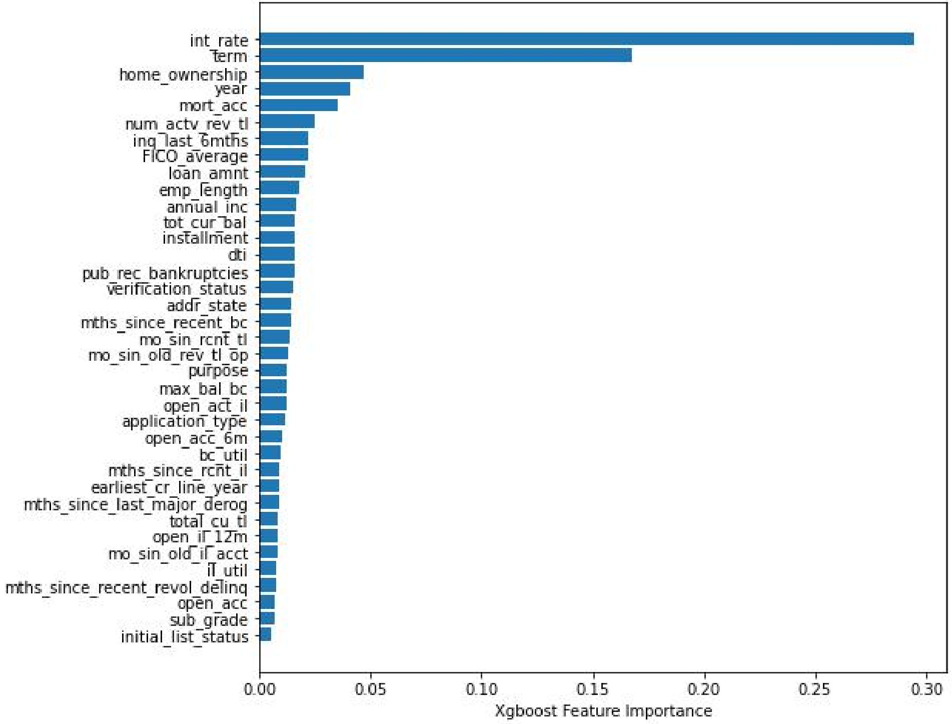


Modelling Results - XGBoost

3. **XGBoost:** ROC AUC Score: 64.55%







XGBoost discards correlated variable when breaking down trees further

However, random forest builds its tree using random selection of features. Since these models behave differently we can see differences in the feature importance

Next steps

1. Perform regression analysis on the interest rates to understand the factors influencing the interest rate of a borrower. This may help lenders to forecast a interest rate based on the important features & evaluate if the actual interest rate is attractive or not
2. Perform hyperparameter tuning more rigorously on the current classification models to obtain better results

Conclusion

1. In this analysis, we understood the various parameters that affect a loan default. Features like Interest Rate, loan term, dti, grades, etc which is also evident from the EDA
2. Analysis also benchmarks different classification algorithms & their feature importances