Title: Loan Default Analysis & Prediction

Authors: Anshul Rao, Kashish Jain, Nikita Demidov, Rahul Pandey

Summary:

P2P(peer-to-peer) lending is the practice of lending money to businesses or individuals via online services that match investors with borrowers.[1] Financial transactions are defined as P2P when they bypass conventional intermediaries by directly connecting the borrower and the lender. Lending Club is one of the world's leading such P2P lending platforms that provides a marketplace to the borrowers & lenders to provide a better experience & help investors earn attractive risk-adjusted returns. However, these lending practices bring greater risk of borrower defaulting & not repaying the loan with interest. In this project report, we'll highlight the major factors contributing to loan defaults & enable investors to make informed decisions regarding the same. We'll be using Lending Club's loan data[2] which consists of over 3M loan applications with a default rate ('Charged Off') of ~20% & with over 140+ predictors. Predictor variable includes customer's credit background, demography, loan specifications, borrower's employment history, etc.

We started by understanding the predictors and how they impact loan default performance. After thorough analysis and discussion we built binary classification models to predict whether a given loan will be fully paid or not by the borrower using Logistic Regression, Random Forest, XGBoost and SVM. Since falsely predicting a loan default is as important as correctly predicting a loan default we will be using AUC ROC as our evaluation metric that evaluates the model's capability of distinguishing between both the classes. As the data is highly imbalanced, we performed under-sampling on our dataset while training on classification model.

Methods:

1. Data Cleaning: We first started by handling the missing values. If the feature had more than 95 percent of NULLs then we dropped the feature, otherwise, we either filled the values with 0, -1, mean or median based on our understanding of the variables. For example, for missing DTI (debt-to-income) ratio it would be unfair to fill the missing values with zeros.
2. EDA: Our next step involved thorough analysis of the features, which involved both univariate and multivariate analysis. We also looked at the features with respect to loan default to get a better understanding of how the features impacted the probability of defaulting. EDA has the below broad categories:-
    2.1. Data Summary
    2.2. Distribution of X variables (Predictors) - Univariate Analysis
    2.3. Relationship between Predictors - Multivariate Analysis
    2.4. Relationship between the Y and X variable (Response and Predictors)
3. Feature Selection: Since the dataset had more than 140+ variables, we spent time narrowing them down to avoid noise and keep only the most relevant ones before we begin modeling. Feature selection includes the following steps:-
    3.1. Deciding on dropping sparse features (having more than 80 percent of 0 values).
    3.2. Deciding on dropping categorical features having all or more than 90 percent of values from the same category.
    3.3. Dropping variables after the loan was approved since we intend to predict default before loan is granted.
    3.4. Reduce multicollinearity by dropping variables with high correlation (>0.7).
    3.5. Dropping irrelevant features like loan id, etc.
4. Transformations and Feature Engineering:
    4.1. The categorical values were either encoded using one-hot encoding or label encoding. For features like *add_state* (US states) it was a better option to use label encoding because otherwise the number of variables would more than double with one-hot encoding.
    4.2. Log transformation was done on very right-skewed predictors like *annual_inc* (annual income) and scaling was done based on the model's requirement.
    4.3. New features like fico = average fico score using low and high fico score and monthly_load = percentage of income that goes in installment every month were also generated.
5. To handle class imbalance, we performed random under-sampling on the non-default observations to evenly distribute the default vs non-default observations.
6. Lastly, we performed classification modelling on the Fully Paid vs Charged Off (Default) using Logistic Regression, Random Forests, XGBoost and SVM with AUC ROC as an evaluation metric.

Results:

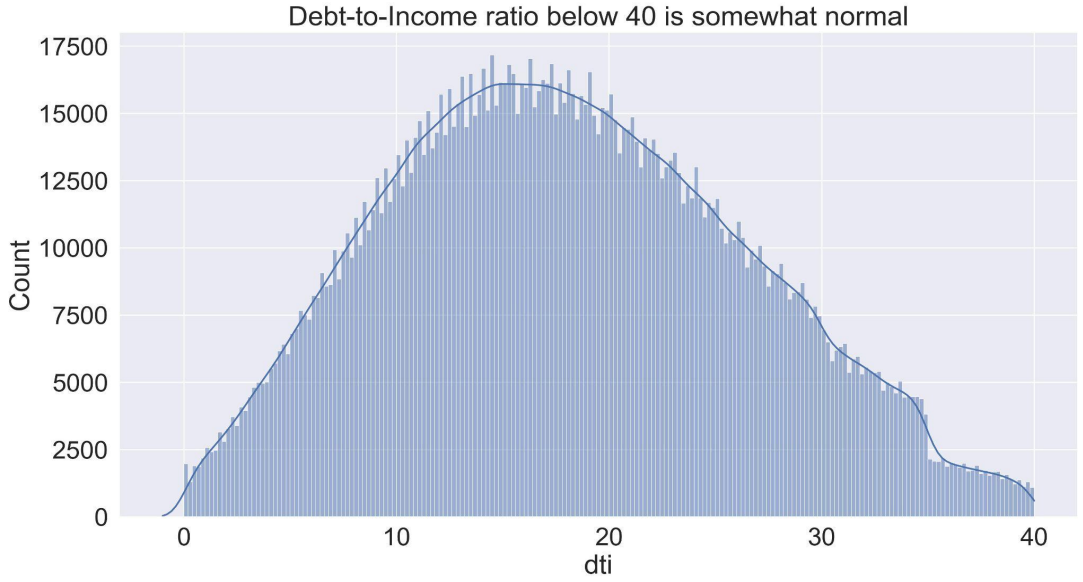
EDA

1. Data Summary

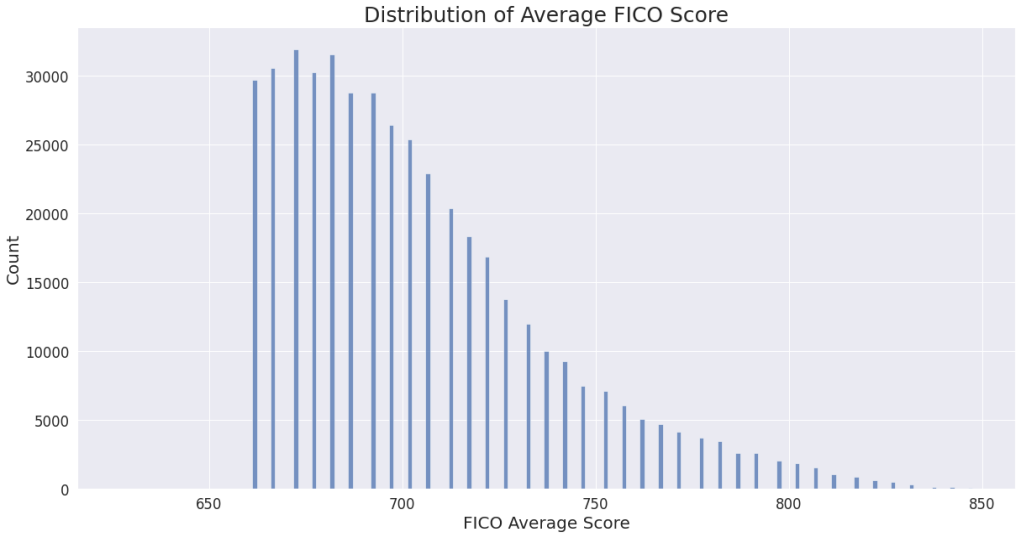
Number of Observations (Completed Loan Applications)	1860764
Number of Predictors	141
Number of Defaults	19.51%
Number of Categorical Columns	35
Number of Numeric Columns	106

2. Distribution of X variables (Predictors) - Univariate Analysis

2.1. DTI values below 40 seem to be somewhat normally distributed with the majority of values falling between 5 and 30.

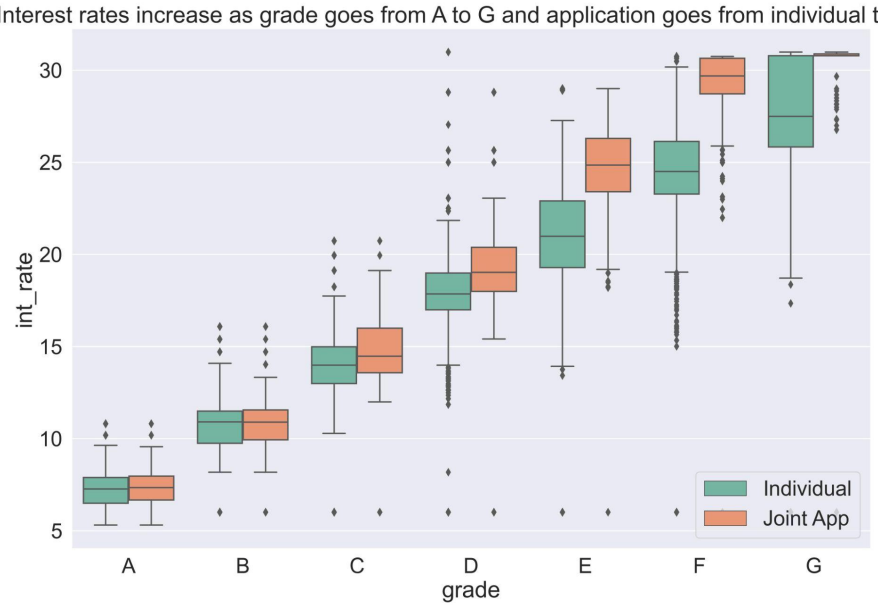


2.2. Average FICO score is right-skewed.

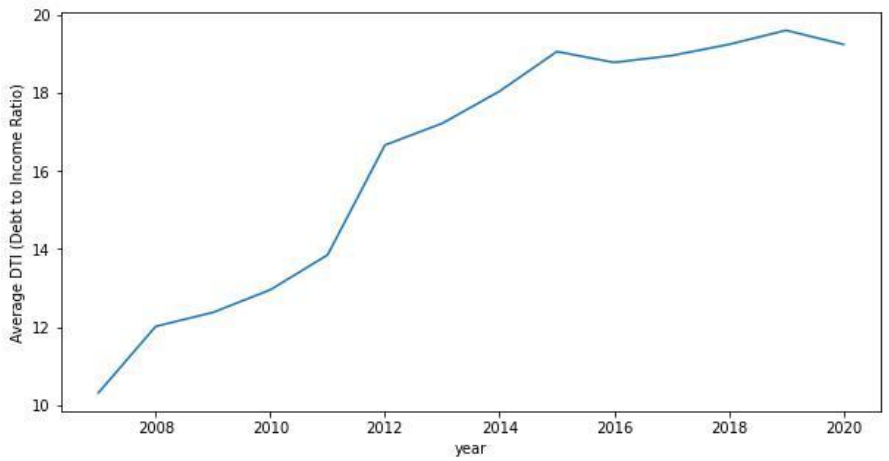


3. Distribution of X variables (Predictors) - Multivariate Analysis

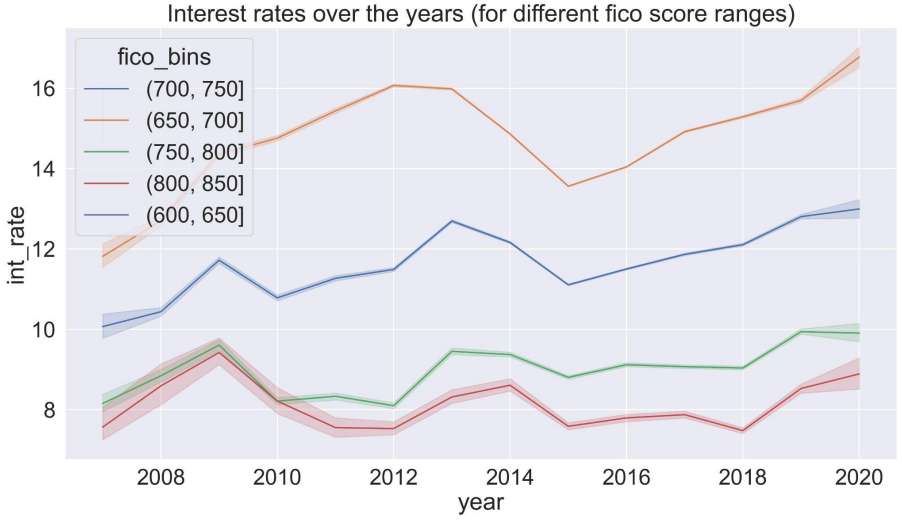
3.1. Interest rates increase as grade goes from A to G and also individual applications have lower interest rates than joint applications and the gap increases as we go from grade A to grade G.



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



3.3. Interest rates are higher for borrowers with low fico score and lower for borrowers with high fico score and while they (interest rates) increased eventually between 2007 and 2020, the increase has not been linear.

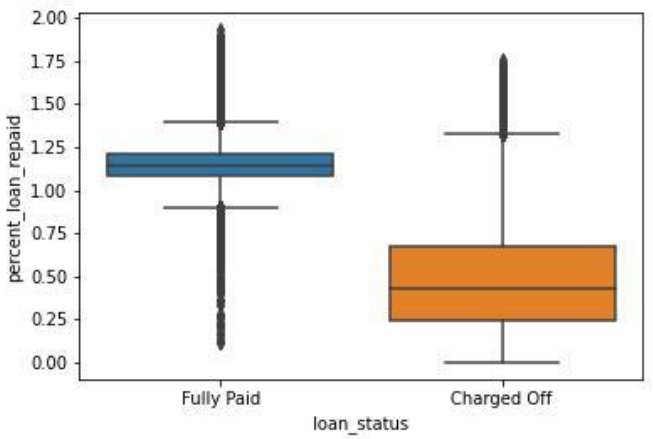


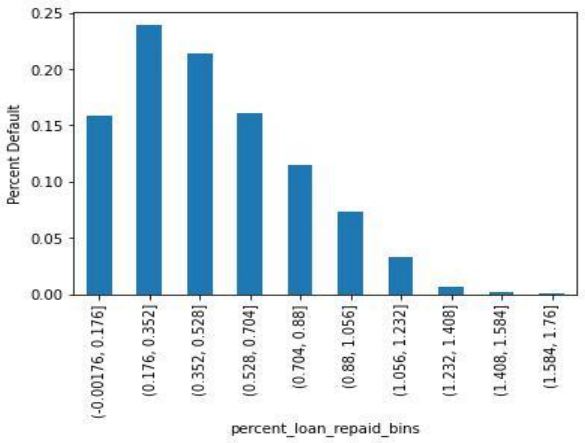
4. Relationship between the Y and X variables (Response and Predictors)

4.1. Interest rates increase as grade goes from A to G

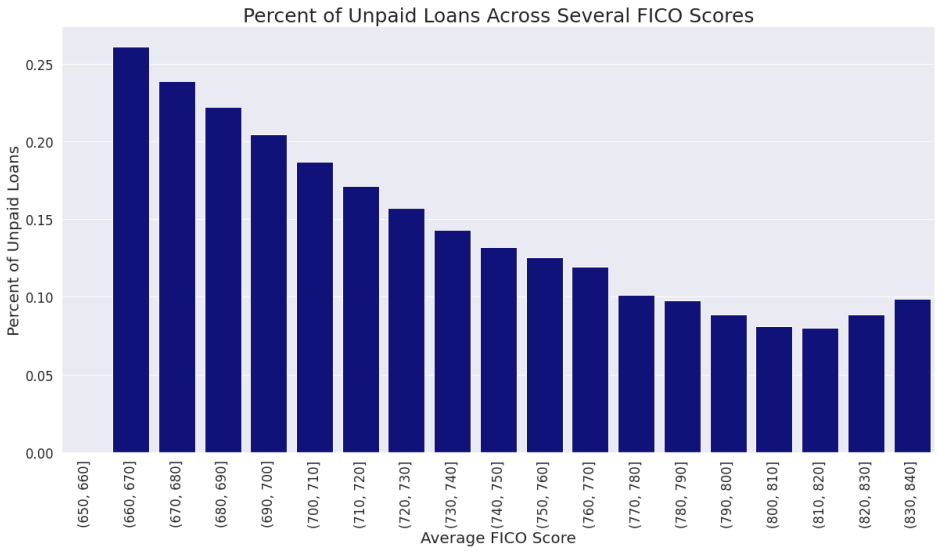
Loans graded A have good repayment chances and it slowly decreases as the grade approches F/G.

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

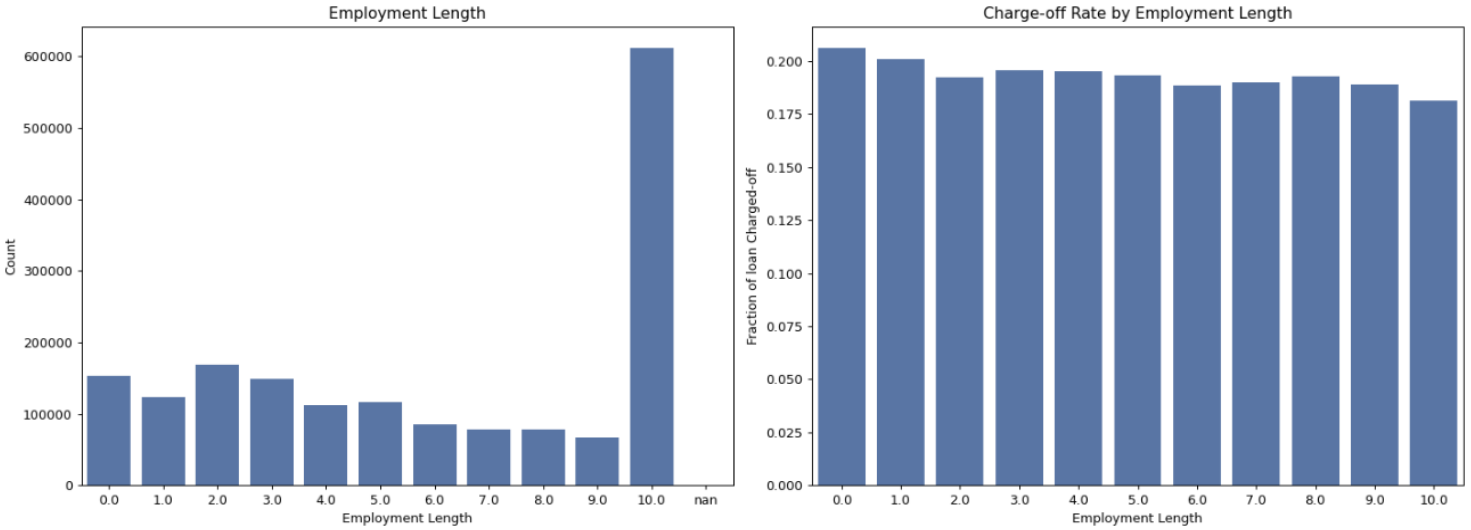




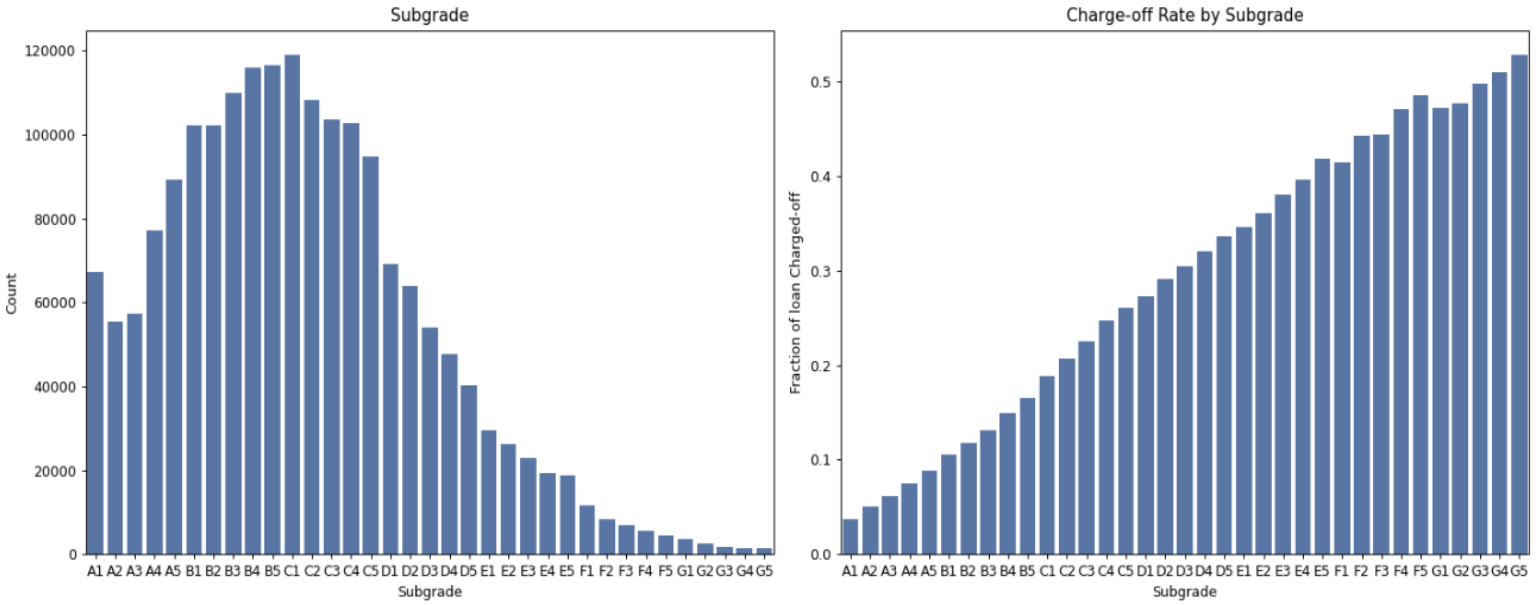
4.3. The lower the FICO score, the higher the proportion of unpaid loans.



4.4. The charge-off ratio is quite uniform across the length of the employment years for instance,.10+ years employment length applicants doesn't show different loan repayment tendency compared to < 2 years employed applicant

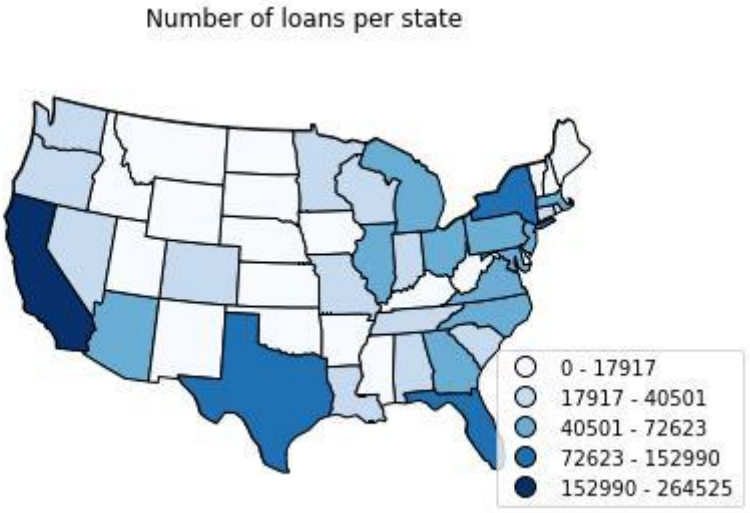


4.5. The charge-off ratio shows a uniform trend which increases from A1 to G5.It reflects that the applicant grading by the lending club is quite accurate.

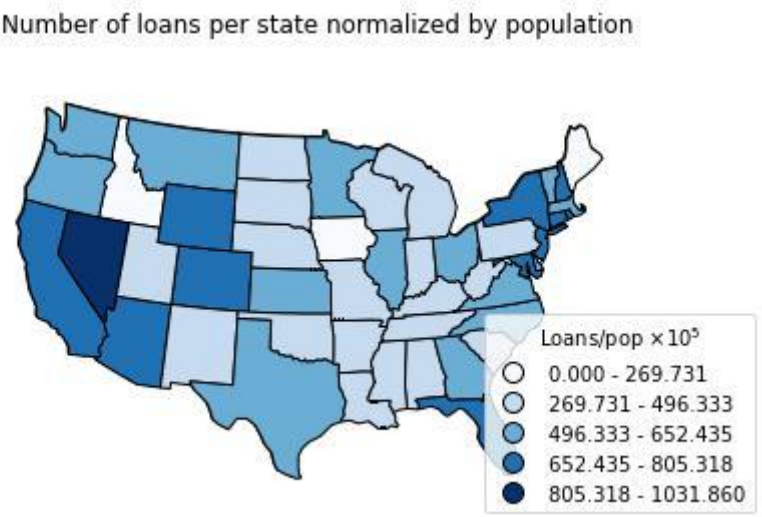


Analysis across US states

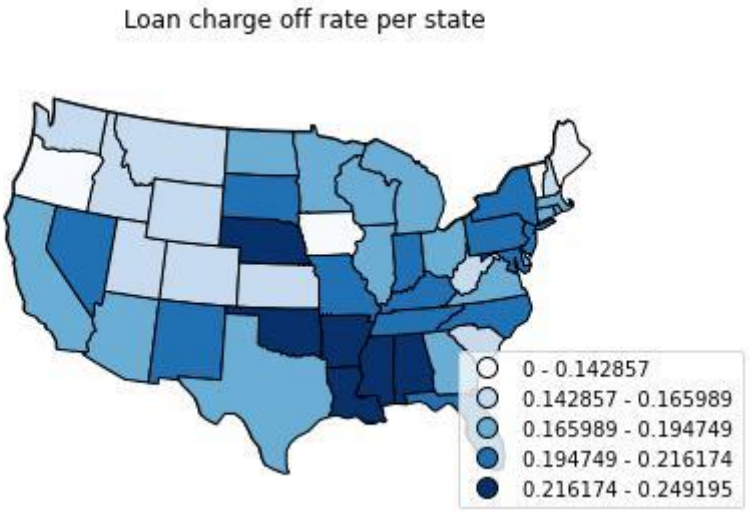
California shows highest number of loan applicants



Number of loans normalised by population shows the highest ratio in Nevada.

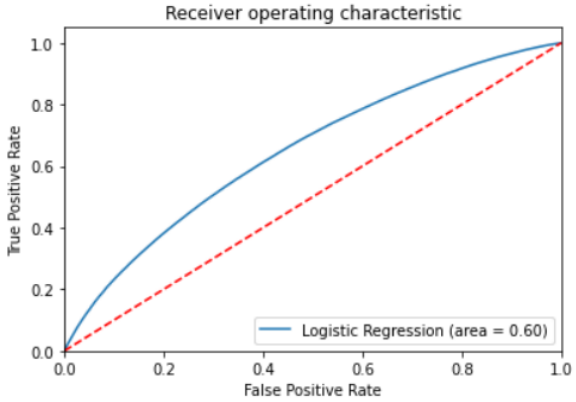


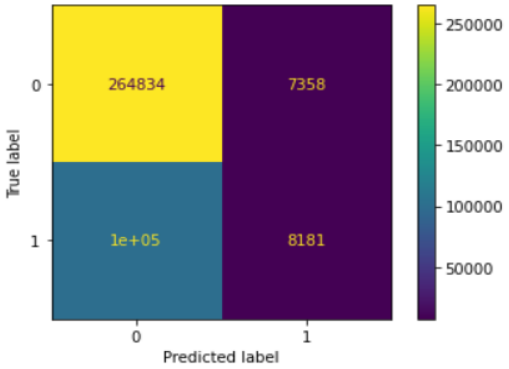
Charge-Off loan percentage is higher in the south-east states i.e. Arkansas, Mississippi, Louisiana, Alabama, Georgia & one of the central state of US i.e. Nebraska.



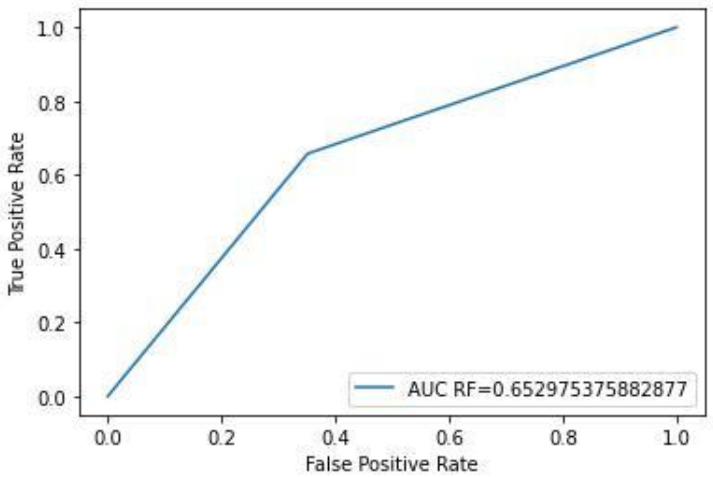
Modeling Results:

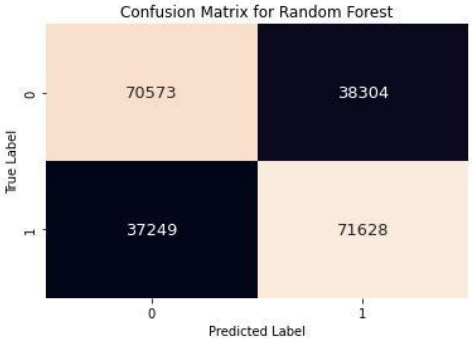
1. **Logistic Regression:** ROC AUC Score: 61.00%



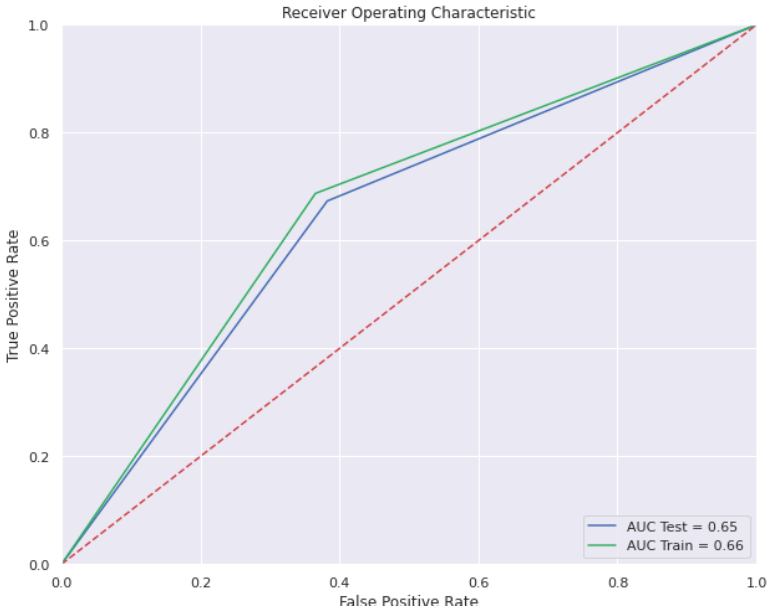


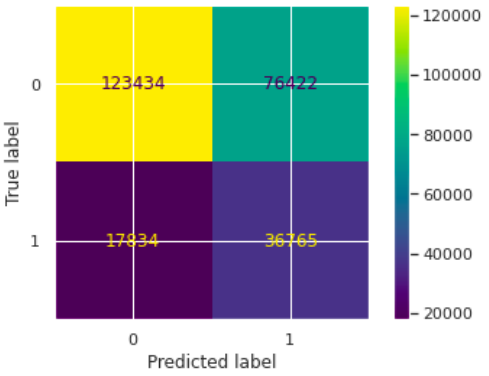
2. **Random Forest:** ROC AUC Score: 65.29%





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





4. **SVM** ROC AUC Mean Score (5 Fold Cross Validation): 67.43%

Discussion:

In our analysis, we found various interesting results like the effect debt-to-income ratio, fico score and interest rates had on default rate. We also came to know that features like grades/sub-grades, which are assigned by the Lending Club, showed a very close and accurate picture of probability of loan default.

In the current version, we understood the important features contributing to a loan default. Furthermore, we can perform a regression analysis on the interest rates of the approved loans to understand the important factors that help decide the borrower's interest rate. This helps lenders to forecast an interest rate based on the important factors & evaluate whether the actual interest rate is attractive or not.

While modeling, some of the processes like tuning of hyperparameters was both time and computationally extensive and needed more resources so we can explore that in depth and see how it can be done more optimally.

Statement of contributions:

Anshul Rao	**Kashish Jain**	**Nikita Demidov**	**Rahul Pandey**
Worked in collaboration with team and finalized handling of missing data for one-fourth of features.	Created Google sheets for feature importances & missing value treatment that helped coordinate work among teammates	Worked in collaboration with team and finalized handling of missing data for one-fourth of features.	Worked in collaboration with team and finalized handling of missing data for one-fourth of features.
Worked independently on EDA and came up with results.	Performed EDA on DTI trend, default-% vs loan repaid & understanding parameters across US states	Worked independently on EDA and came up with results.	Worked independently on EDA and came up with results.
Worked in collaboration with the team on initial feature selection and transformations to be adopted for them.	Worked in collaboration with the team on initial feature selection and transformations to be adopted for them	Worked in collaboration with the team on initial feature selection and transformations to be adopted for them.	Worked in collaboration with the team on initial feature selection and transformations to be adopted for them.
Worked independently on XGBoost modelling technique for classification.	Performed Random Forest classification while filtering variables from logistic regression's p-value.	Worked independently on SVM modelling technique for classification.	Worked independently on Logistic Regression modelling technique for classification.

References:

1. Wikipedia:Peer-to-peer lending, https://en.wikipedia.org/wiki/Peer-to-peer_lending
2. Kaggle:Lending Club 2007-2020Q3, https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1
3. https://cs229.stanford.edu/proj2018/report/69.pdf
4. https://www.researchgate.net/publication/340395124_Project_Lending_Club_Data_Analysis
5. https://www.lendacademy.com/lendingclub-closing-down-their-platform-for-retail-investors/

Appendix:

Handling Missing Values:

- https://docs.google.com/spreadsheets/d/1d0taR-dpsgiWGYGwrvQKiuBPWI3Z8IVcDjPIU1GUmv8/edit#gid=342959862

Feature Selection:

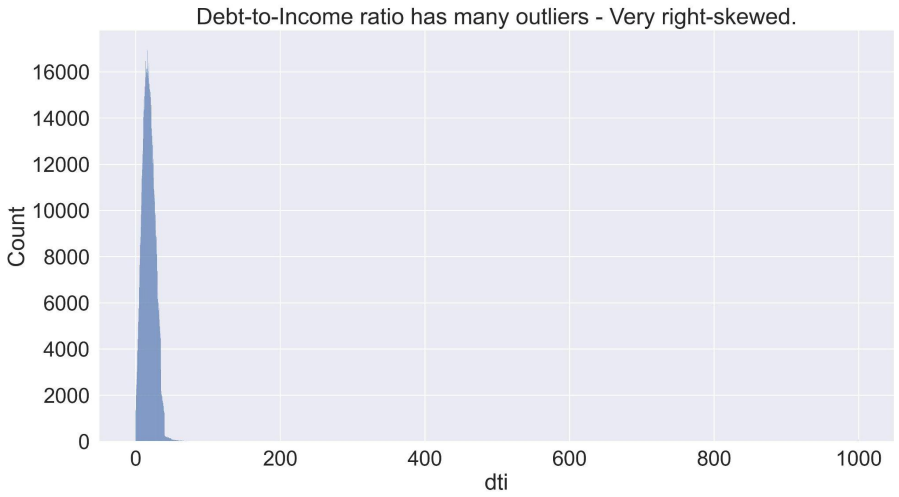
- https://docs.google.com/spreadsheets/d/1d0taR-dpsgiWGYGwrvQKiuBPWI3Z8IVcDjPIU1GUmv8/edit#gid=1642151515
- https://docs.google.com/spreadsheets/d/1q4T6MC4Og5_JiXUOiwRTDq9nE6BjOlbVxREcwg2bVU4/edit#gid=1642151515

EDA and Modeling:

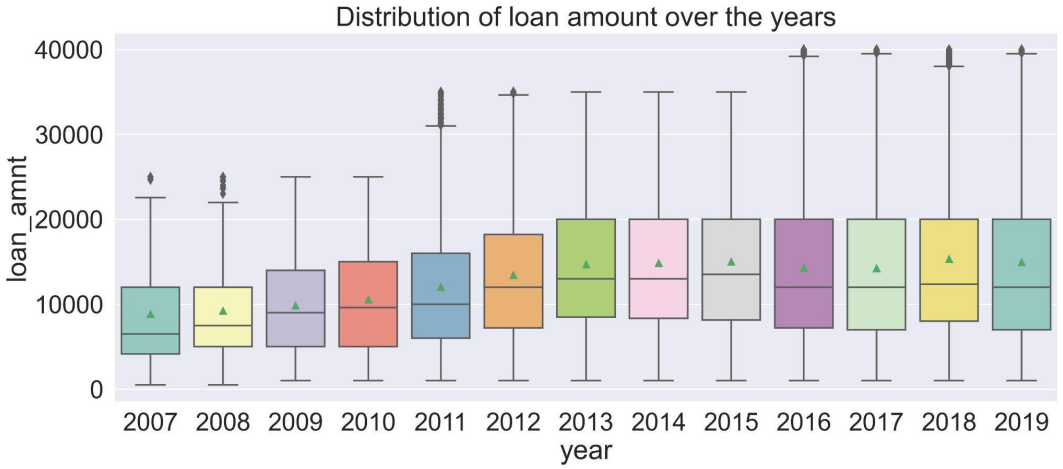
- https://www.kaggle.com/code/anshulrao/loan-default-analysis-and-prediction-python
- https://www.kaggle.com/code/jkashish18/lending-club-python
- Lending_Club_Python f38de7 | Kaggle
- https://colab.research.google.com/drive/1HNFFeIVETapiGCTSIozHKX28Rr3FBnUM?usp=sharing

Supplementary Figures:

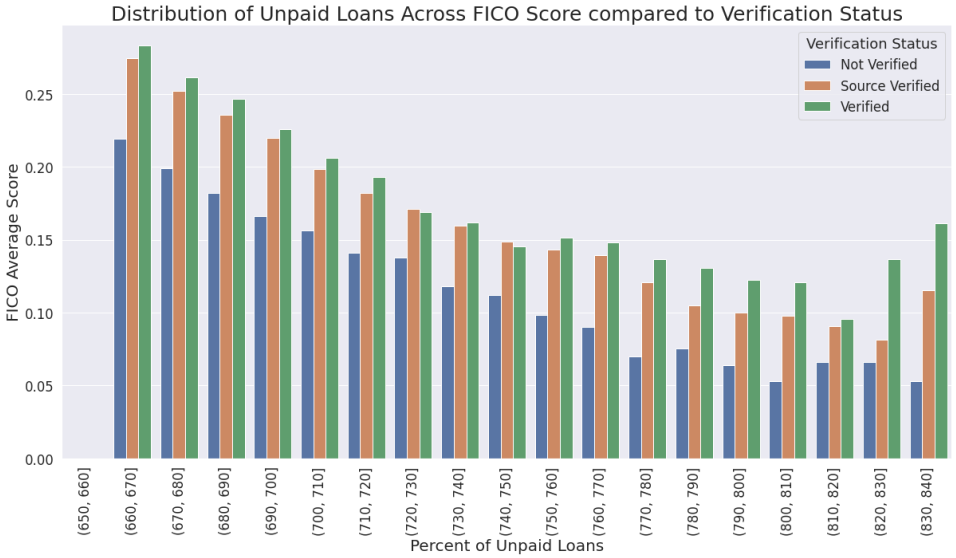
- The debt-to-income (DTI) ratio is very right skewed.



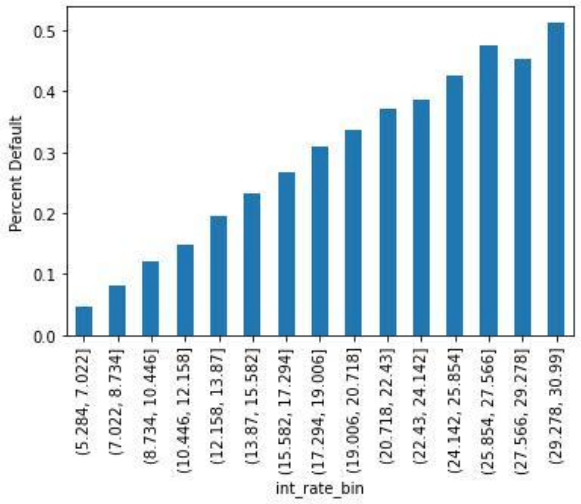
- The average loan amount steadily increased from 2007 to 2013 and then it stayed consistent till 2019.



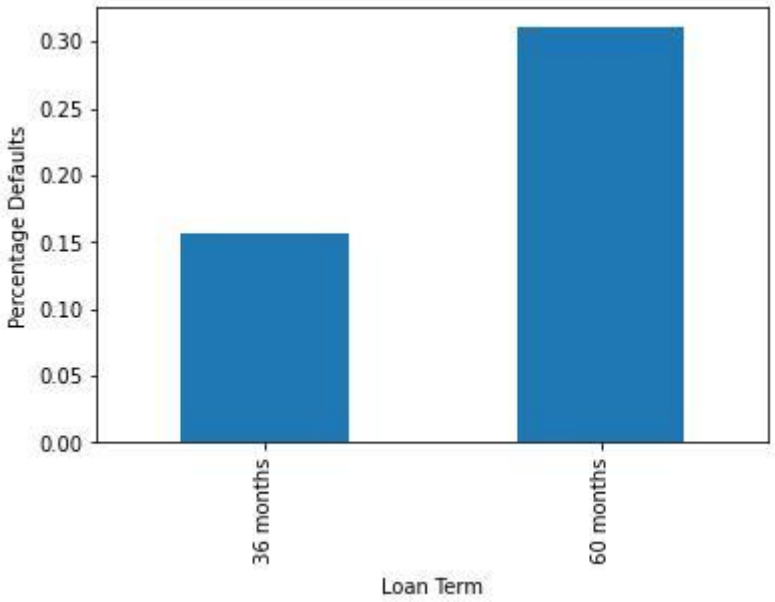
- Verified and source verified income status have constantly higher percent of unpaid loans across all of the FICO average score



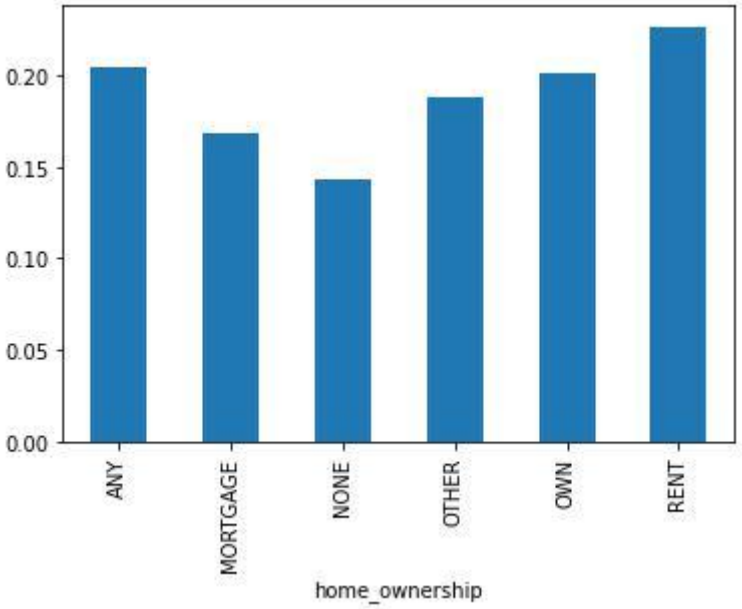
- Interest Rate v/s Default Percentage



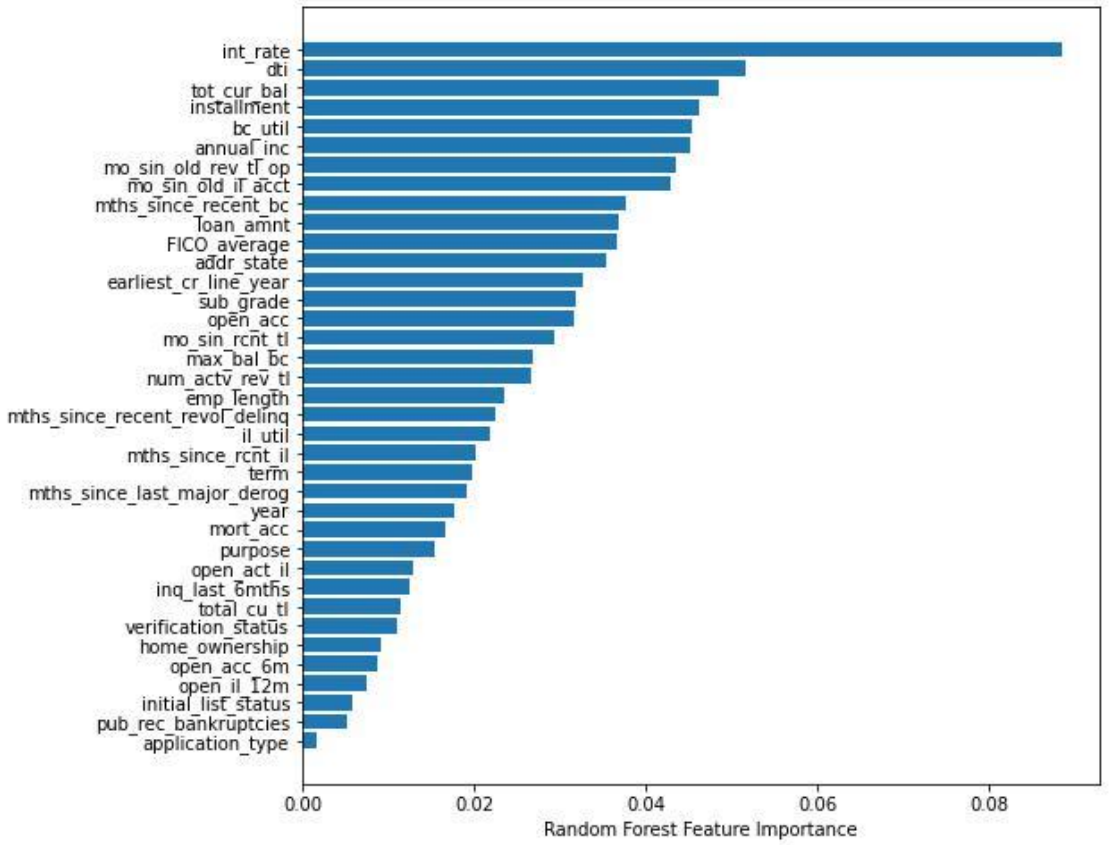
- Loan Term v/s Default Percentage



- Home Ownership v/s Default Percentage



- Random Forest Feature Importance



- XGBoost Feature Importance