Loan Approval Rates and Influencing Factors in India

Project Contributors: Shay Garcia, Cris Mariano, Giovanna Hayes, Richard Tango

Report Prepared By: Giovanna Hayes

September/October 2020

Contents

i. Problem Statement

This project was inspired by an interest in personal credit and trying to answer the question what factors influence loan approval rates? Is it macroeconomic factors, demographic factors, the loan attributes themselves or a combination?

One way to answer this question is the analyze a data set of loan applications, inclusive of information regarding the loan applicants and the loan application outcomes. It will also be beneficial to analyze a data set with macroeconomic performance indicators as well.

There are multiple groups of stakeholders who would be interested in the findings of this report. The primary customers of our findings would be financial institutions. These findings could be used in the creation of consumer products servicing these credit vehicles. Our secondary customers for our findings would be credit applicants who are interested in understanding what factors could impact their chances of having their loan approved. We feel this information could be valuable to this group as well.

ii. Data Collection and Wrangling Summary

Data Collection

We collected the data from Kaggle and Quandl:

https://www.kaggle.com/mishra5001/credit-card?select=application_data.csv

https://www.quandl.com/data/BSE/BSE500-BSE-500

This dataset from Kaggle consists of over 308K loan application records from IIT Bangalore, India from the year 2019. We use the application_data.csv and previous_application.csv files and upload them into Pandas data frames. This dataset from Quandl consists of open and close prices from the years 2019 and 2018 from the Bombay Stock Exchange's BSE500. This ETF encompasses the largest 500 companies on the Bombay Stock Exchange and is a strong indicator of the overall market performance.

Irrelevant Data and Missing Values

For the purpose of this project we decided to remove loan attribute data related to the timing of when the loan was submitted or previously submitted. We also removed data related to the credit bureau rating.

Upon investigation of the loan data frame we also discover that the Occupation Type had 91,240 missing values, representing 30% of the total count of Occupation Type values. We decided to assign the value, Laborer, to these empty values given Laborer is the largest Occupation Type identified within the dataset across all age bins.

ii. Exploratory Data Analysis

Loan Approval Rates in 2019 was 77%. Approved loans include unused offers and approved loan statues.

		Number of Loans	Percentage
0	Approved	224045	77
1	Unapproved	67012	23
2	Total	291057	100

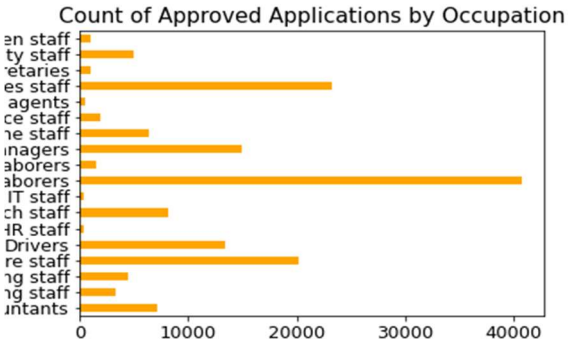
The overall approval rate was 77 percent

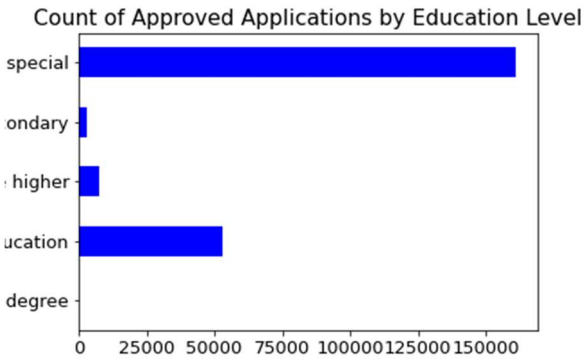
<u>Demographic and Macroeconomic Factors</u>

Impact of Education and Occupation on Approval Rates

Most applicants who were approved had a secondary or special secondary educational background.

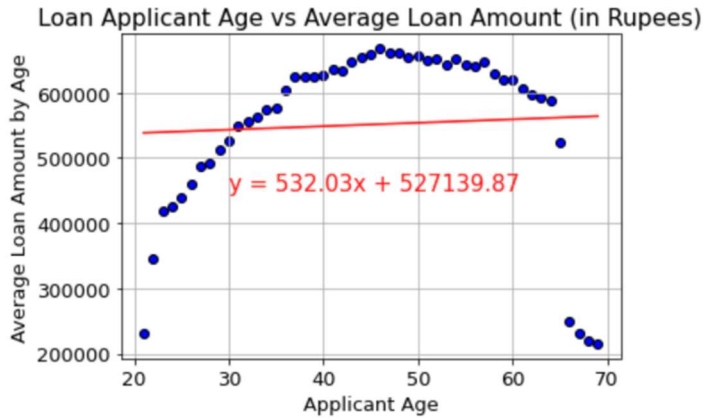
There was a total of 160,874 approved apps for those with secondary/special secondary education level. Most applicants who were approved were Laborers. 40,726 laborers had their loan application approved.

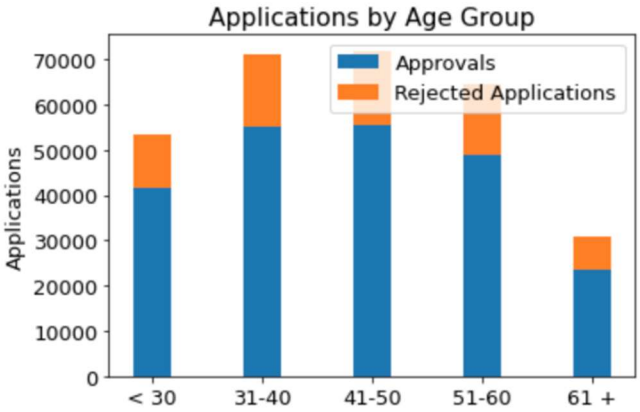




Impact of Age on Approval Rates

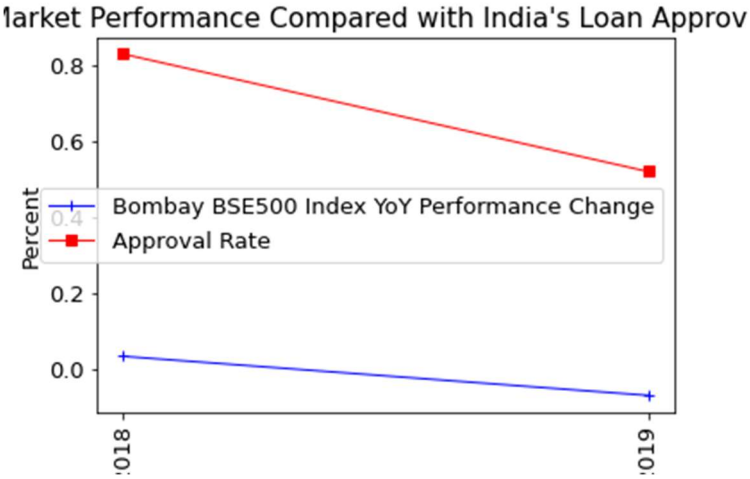
There is no direct correlation between age and approval rates. Average loan credit application amounts increased with age. At approximately age 55, the average credit application amount began to decrease as age started to increase.





Impact of Stock Market on Approval Rates

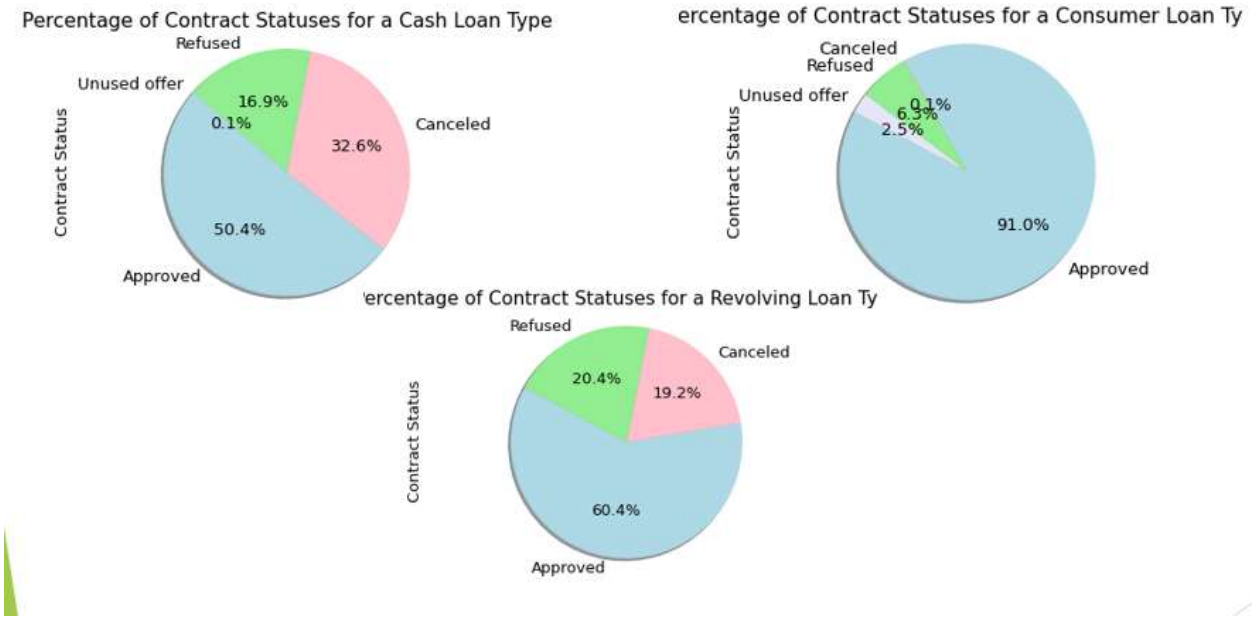
While there is direct correlation between market performance and approval rates there is not enough data to support market performance influences loans approval rates.



Loan Attribute Factors

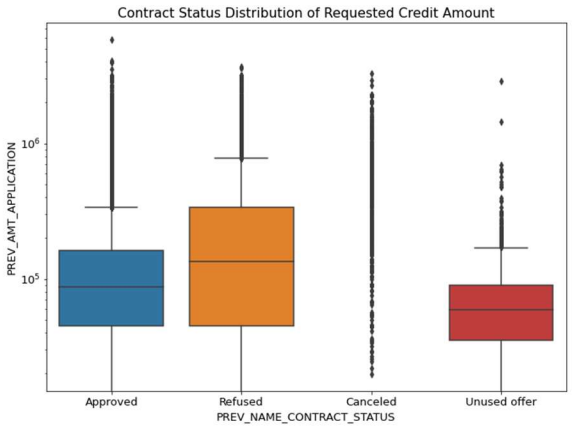
Impact of Type of Requested Loan on Approval Rates

Consumer loans had a 94% approval rate on average which was a significantly higher rate of approval when compared to cash loans and revolving credit loan types.



Impact of Type of Requested Loan on Approval Rates

The requested credit amounts are all normally distributed excluding the unused offer status because of the low count of data points. The median requested credit amount for approved contracts is 87,714 rupees. The median requested credit amount for refused contracts is 135,000 rupees. It seems that there are a lot of outliers within the Approved status box plot.



T-Tests

After performing the Exploratory Data Analysis (EDA) on the data we came to some preliminary conclusions on our findings.

Below we will perform t-tests to confirm which of our EDA findings are statistically significant. More specifically:

1) We will perform a t-test on two independent samples: the approved loans with a specific attribute and the approved without it

2) We will test the null hypothesis $H0$ that the approval rate for the two samples is identical.

2) The alternative hypothesis Ha would be that the approval rates are different

3) If the t-test results are statistically significant (e.g. p-value > a,a = 0.05), then we will reject the $H0$ and accept the Ha .

Age

EDA: There is no direct correlation between loan approval rates and the loan applicant's age.

The p-value of 1.455248556248674e-44 is very small, so we can reject the null hypothesis and confirm that age contribute to a higher approval rate. This does not necessarily align with our original EDA findings noted directly above.

Credit Amount

EDA: The median requested credit amount for approved loans was 88K Rupees. The median requested credit amount for loans that were not approved was 135K Rupees.

The p-value of 0.0 is very small, so we can reject the null hypothesis and confirm that the amount of credit requested on the loan application correlates to higher approval rates. This aligns with our original EDA findings noted directly above.

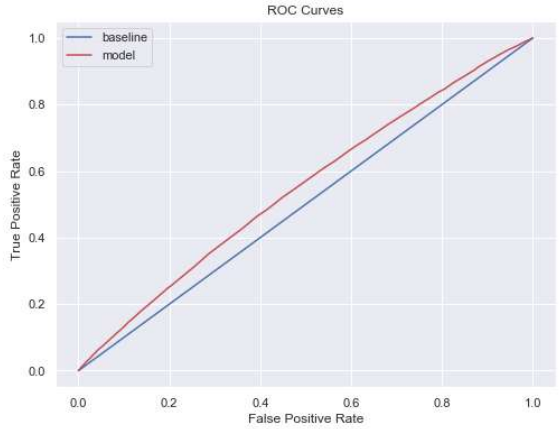
iv. In-Depth Analysis using Machine Learning

Predictive Modeling: Choosing the Best Model

We will try Logistic Regression, Decision Tree and Random Forest models. For each of the models we will be using the sklearn package.

Our random forest has produced the best f-1 score (0.771) indicating it is potentially the most accurate. We will plot the ROC score against the baseline to validate the accuracy of the model.

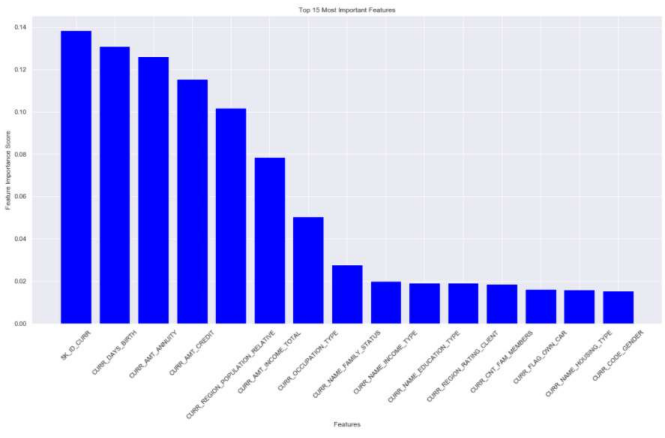
Our ROC Curve is very much in line with the baseline prediction curve further validating the accuracy of the Random Forest model



Findings

We review top 15 most important features of the models by using sklearn featureimportances to help us understand the average impact each factor has on the model output.

Excluding the loan applicant ID we can see that the applicant's age (seen in the metric CURR_DAYS_BIRTH), the amount of the loan (as see through the metrics CURR_AMT_ANNUITY/CREDIT), surprisingly where they are from and finally the applicant's annual income are key factors influencing the prediction outcomes of the random forest model.



v. Recommendations

Credit applicants who are more advanced in age and who hold secondary degrees are more likely to get approved for loans in India based on the 2019 dataset but not for all loans

types. These applicants are more likely to be approved for consumer loans rather than cash and revolving credit loans. Regardless of all influencing factors for loan applicants, requested credit amounts will always heavily influence the probability of a loan being approved. It is recommended to note this above all else.

vi. Ideas for Further Research

While there is a correlation between macroeconomic factors such as the stock market performance and approval rates, we suggest additional in-depth analysis to confirm a conclusion that strong market performance could indicate a higher probability of loans being approved. We would also recommend performing a similar analysis on a dataset with loan applications from the United States to compare and contrast the average impact similar factors have on the model outputs.