Interest Rate Category Prediction Management System

Summary:

Banks offer various kinds of accounts and provide loans based on the requirements. Apart from it, there are various activities like investments in market and different funds. Overall, Banking sector has a wide impact on the economy directly and indirectly. There are many banks across the globe that are leveraging Machine Learning and AI in their daily routine and getting benefits out of it. For example, top banks in US like JP Morgan, Wells Fargo, Bank of America, City Bank and US Banks are already using Machine Learning to provide various facilities to customers as well as for risk prevention and detection. Some of the applications include Customer Support, Fraud Detection, Risk Modeling Customer Segmentation and Predictive Analytics.

We see that lenders use various factors such as credit score, annual income, loan amount approved, tenure, debt to income ratio etc. in deciding the interest rates. The process, defined as 'risk-based pricing', uses a sophisticated algorithm that leverages different determining factors of a loan applicant. Selection of significant factors will help develop a prediction algorithm which can estimate loan interest rates based on clients' information. On one hand, knowing the factors will help consumers and borrowers to increase their credit worthiness and place themselves in a better position to negotiate for getting a lower interest rate. On the other hand, this will help lending companies to get an immediate fixed interest rate estimation based on client's information.

The goal of this project is to predict the Interest Rate Category (1,2,3) Low, Medium, High for each Loan applicant. It is a Multi Class Classification Problem and the Interest Rate Category depends upon several features. Primarily, I will be focusing more in High Interest Rate category and Low Interest Rate Category because classifying High Interest Rate Category as Low Interest Rate Category will be potential harm to banks and classifying Low Interest Rate Category as High Interest Rate category will be potential harm to Customers

Data Description:

The Loan Applicant information data set used in this analysis consists information about 164309 loan applicants and their Interest Rate Categories comprising of 14 different features such as Loan Amount Requested, Length Employed, Annual Income, Debt to Income ratio, Number of Open Accounts, Gender etc.

Feature	Description
Loan_ID	A Unique ID for Loan
Loan_Amount_Requested	Listed amount of the loan applied for by the borrower
Length Employed	Employment Length in Years
Home Owner	Home Ownership status provided by borrower during registration
Income Verified	Indicates if income was verified, not verified, or if the income source was verified
Purpose of Loan	A category provided by borrower for loan request
Debt to Income	Ratio calculated using borrowers total monthly debt payments on debt obligations, excluding mortgage and requested loan , divided by borrowers self reported monthly income
Inquiries Last 6 Months	Number of by creditors during the past 6 months
Months Since Delinquency	Number of months since the borrowers last Delinquency
Number of Open Accounts	Number of Open credit lines in borrowers credit file
Total Accounts	Total number of credit lines currently in borrowers credit file
Gender	Gender
Interest Rate	Target Variable: Interest Rate category (1/2/3) of application

Data Preprocessing:

There are many missing values in several features of a dataset and the number of missing values in each feature is given in the below table

Feature	No of missing Values	Description
Length_Employed	7371	Employment Length in Years
Home_Owner	25349	Home_Ownership Status(Rent, Own, Mortgage..)
Annual_Income	25102	Annual Income provided
Months_Since_Delinquency	88379	Number of Months since the borrowers last delinquency

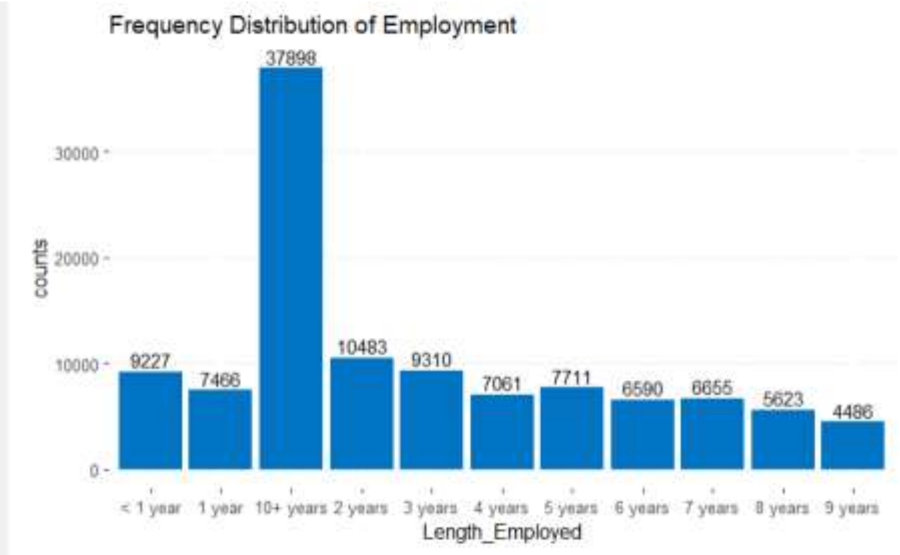
The percentage of missing values is very less in Length Employed i.e. less than 0.05% when compared with actual dataset and the percentage of missing values in Home Owner and Annual Income is also comparatively less when compared with total number of observations in dataset. More than 50% of the observations in the Months Since Delinquency feature is having missing values. In general, we use different Imputation techniques such as Central Imputation, k-nn Imputation etc. to handle missing values and without having strong understanding on domain simply imputing with mean or mode will not produce better results instead it will add more bias in the dataset.

Also, the percentage of observations is less when compared with actual dataset dropping these observations will not impact our analysis. However the number of missing values in Months Since Delinquency is more than 50% But based on my domain understanding Months Since Delinquency plays a major role in deciding the interest rate category because lesser the value of Months Since Delinquency higher the chances of getting Interest Rate higher or sometimes the applicant could not avail loan as well based on background check. However, these missing values can be interpreted in another way i.e. for the applicants who do not have any values for Months Since Delinquency there is a chance that those applicants did not committed any crime till date. So, instead of having this ambiguity I have created a new Feature named Delinquency Status with Yes or No i.e. either person committed crime or not

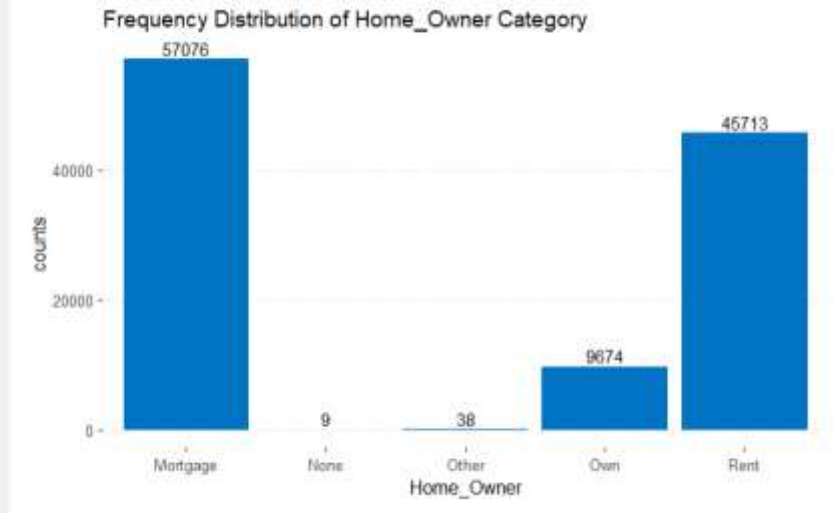
Loan Applicant ID is unique to each applicant and having this feature in our model would not add any significance. So, dropping this feature will not impact our analysis and converted all int or numeric features to numeric and all categoric features to factor levels.

Exploratory Data Analysis:

- Most of the Loan Applicants have been employed for more than 10+ years

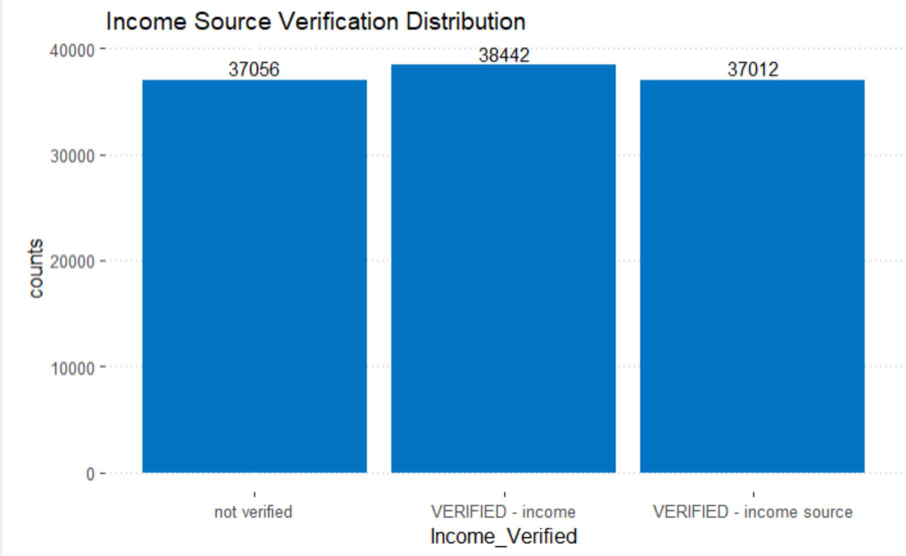


- More than 90% of Loan Applicants Home Owner Category belongs to either Mortgage or Rent

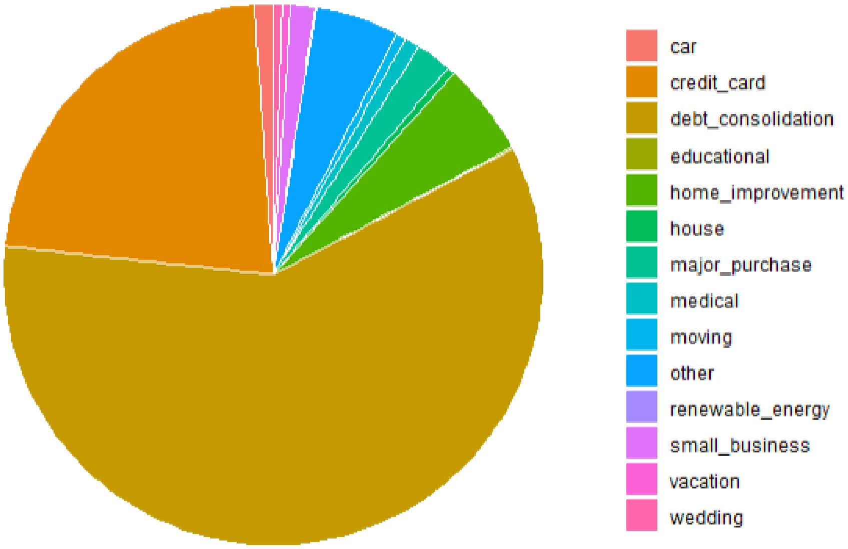


There are very few observations in Other and None that belongs to Home Owner Category say approx. 0.01%. So, either we can drop these observations or else we can create new level by combining these two levels. As part of this project, I combined these two levels and created a new level

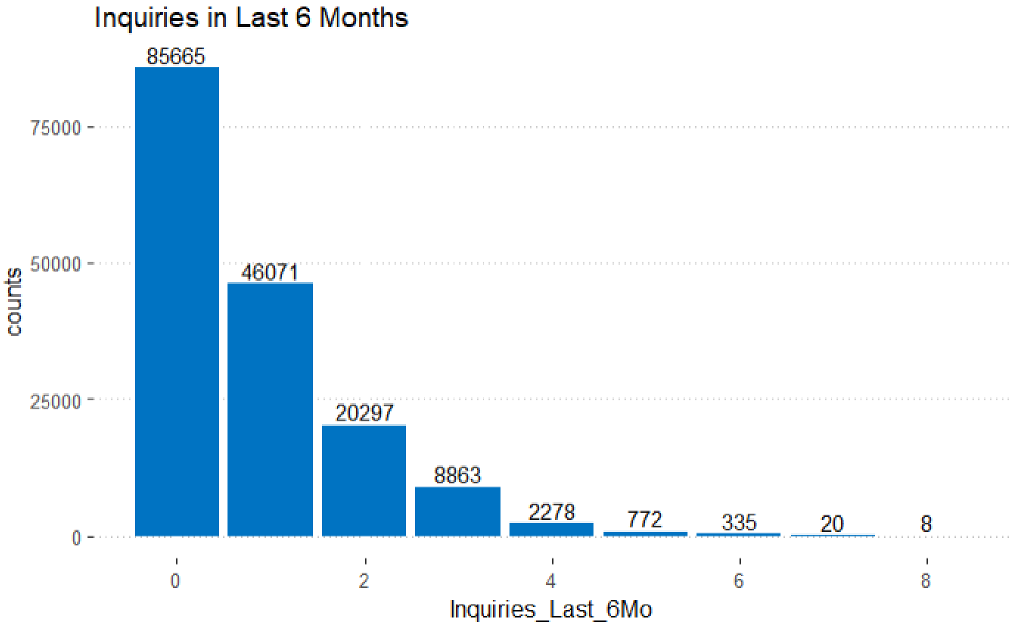
- Majority of the Loan Applicants Income is Verified but still there are significant number of Loan Applicants whose Income is not Verified and there are almost equal number of applicants as that of Income Verified whose Income source is Verified but not the actual Income



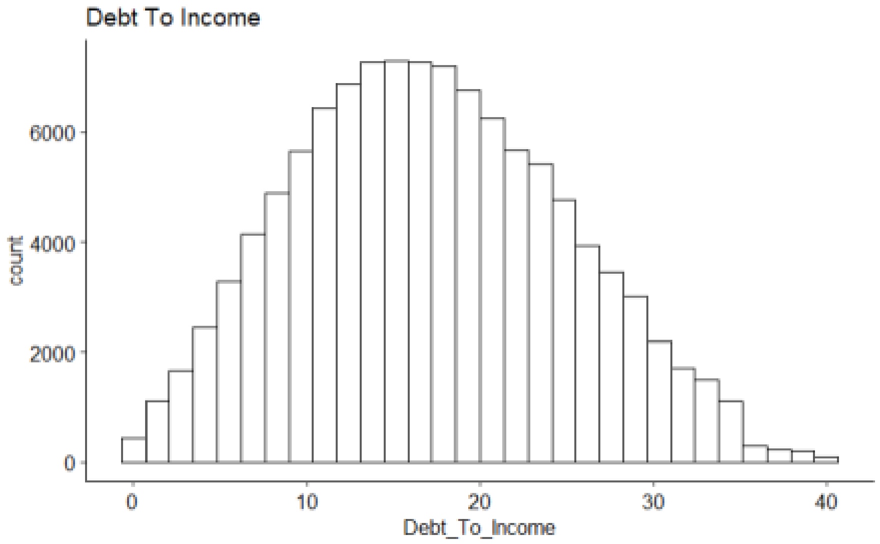
- Majority of Loan Applicants main purpose of taking Loan is either credit card or debt consolidation



- Most of the Loan Applicants did not made any inquiries in the last 6 months but there are few applicants who made 8 inquiries as well in the last 6 months

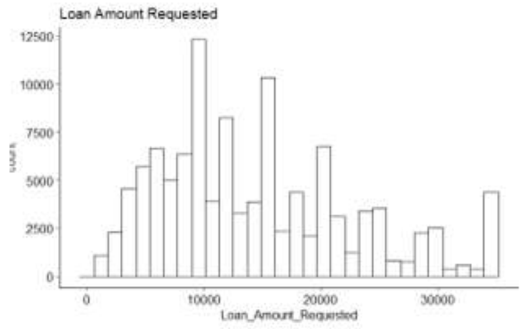


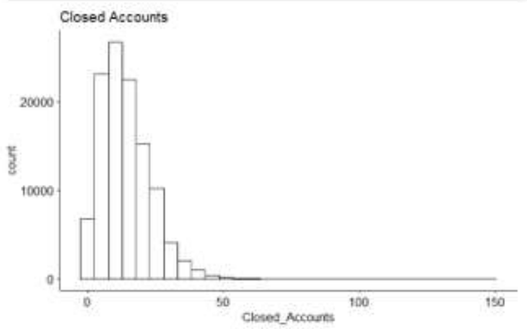
- Debt to Income Ratio is Normally distributed or its distribution is seeming to be approximately Normal Bell-Shaped curve

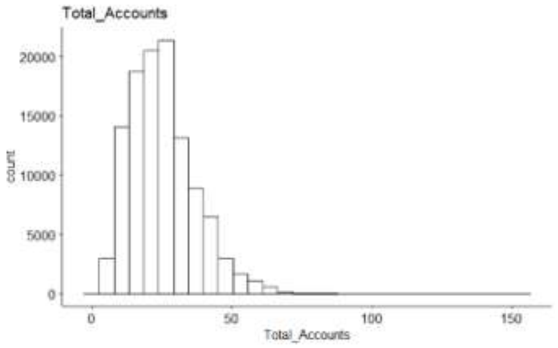


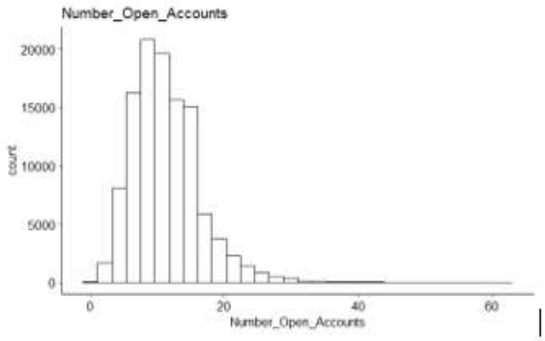
- There is high Skewness and Outliers in Loan Amount Requested, Open Accounts, Closed Accounts, Total Accounts and Annual Income

Before Transformation:

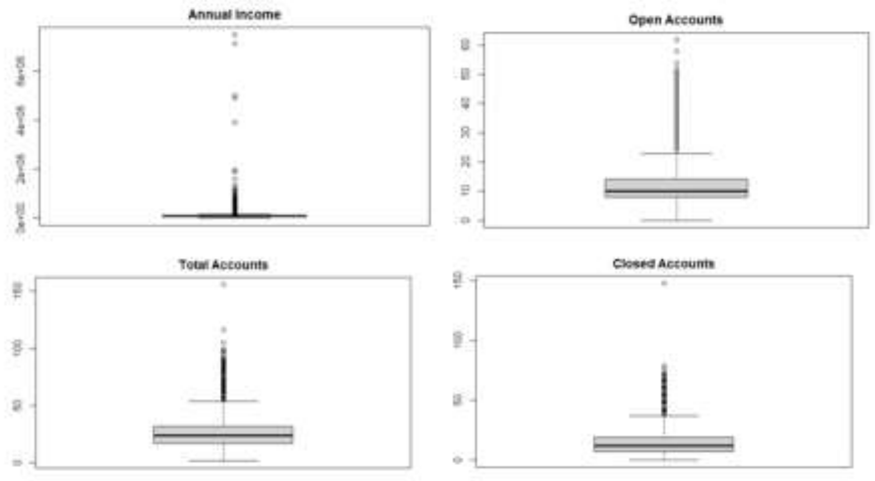






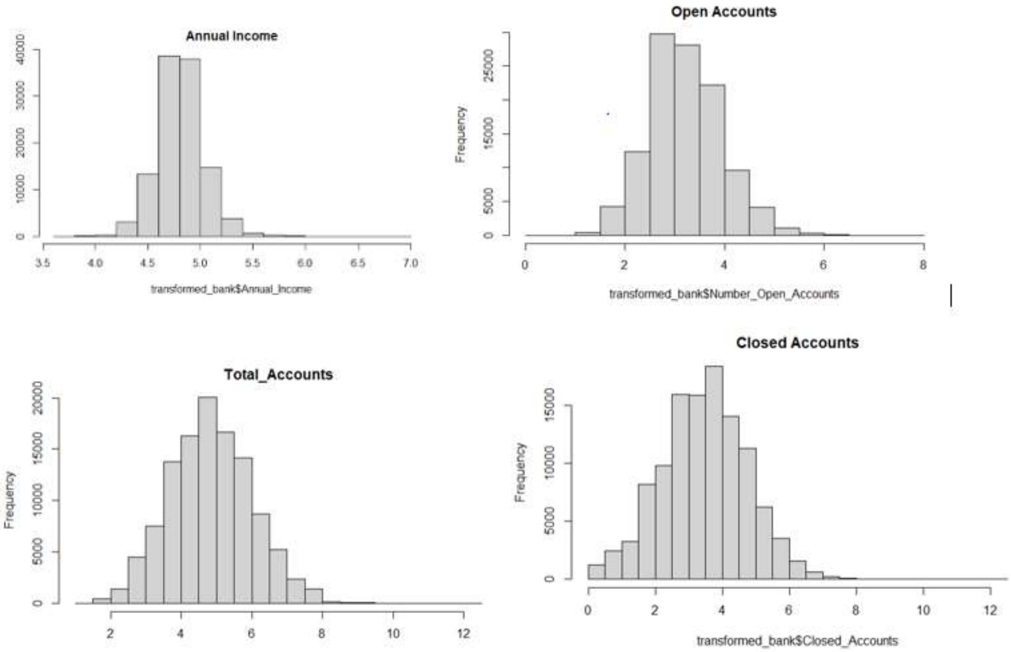


Outliers:

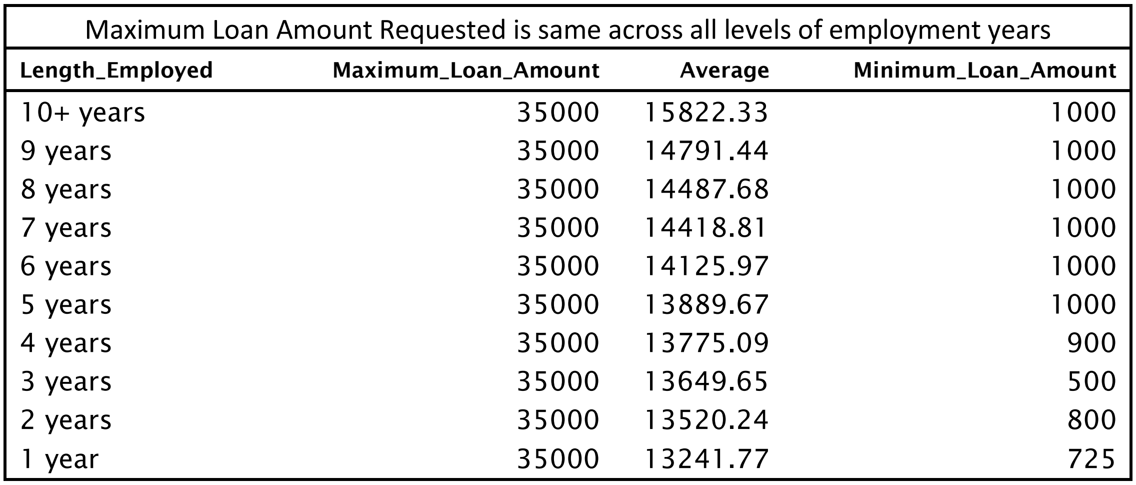


- After appropriate transformations such as square root and logarithmic transformations the distribution becomes Normal or approximately normal

After Transformations:



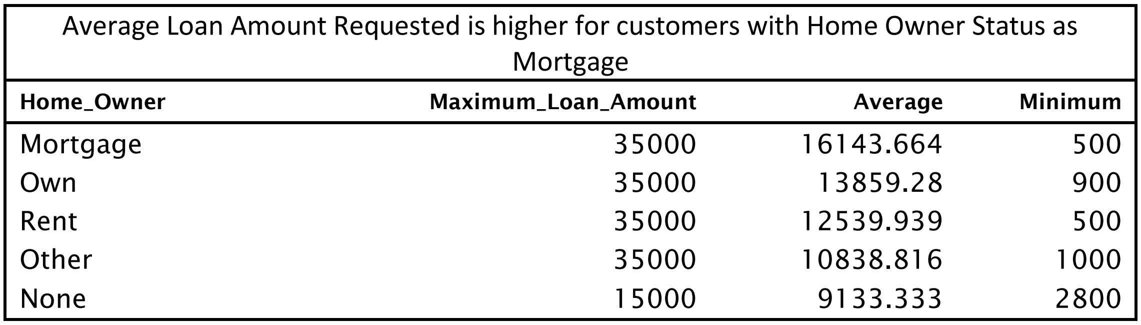
- Maximum Loan Amount Requested across all levels of Employment Years is same



Maximum Loan Amount Requested is same across all levels of employment years

Length_Employed	Maximum_Loan_Amount	Average	Minimum_Loan_Amount
10+ years	35000	15822.33	1000
9 years	35000	14791.44	1000
8 years	35000	14487.68	1000
7 years	35000	14418.81	1000
6 years	35000	14125.97	1000
5 years	35000	13889.67	1000
4 years	35000	13775.09	900
3 years	35000	13649.65	500
2 years	35000	13520.24	800
1 year	35000	13241.77	725

- Average Loan Amount Request is higher for applicants with Home Owner Status as Mortgage



Average Loan Amount Requested is higher for customers with Home Owner Status as Mortgage

Home_Owner	Maximum_Loan_Amount	Average	Minimum
Mortgage	35000	16143.664	500
Own	35000	13859.28	900
Rent	35000	12539.939	500
Other	35000	10838.816	1000
None	15000	9133.333	2800

- Average Loan Amount Requested is higher when the purpose of Loan is Small Businesses and Housing Related

Purpose_Of_Loan	Maximum_Loan_Amount	Average	Minimum
small_business	35000	15644.41	500
house	35000	15294.82	1000
debt_consolidation	35000	15262.15	800
credit_card	35000	14915.53	725
home_improvement	35000	14010.61	900
renewable_energy	35000	11141.05	1000
major_purchase wedding	35000 35000	10496.17 10049.25	1000 1000
other	35000	9762.511	500
medical	35000	9140.375	1000

- Average Loan amount Requested is highest for Interest Rate Category 3 (High) and there is no significant difference in average between Interest Rate Category 1 and Interest Rate Category 2

Average Loan Amount is highest for Interest Rate Category 3(High) and no significant difference between Interest Rate Category 1 and Interest Rate Category 2

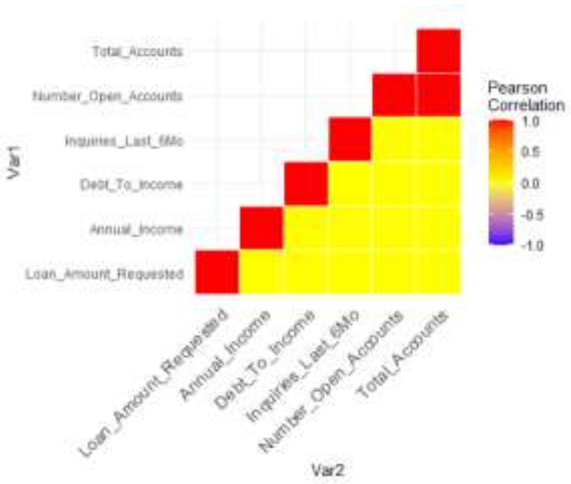
Interest_Rate	Maximum_Loan_Amount	Average	Minimum
3	35000	16214.57	1000
2	35000	13526.8	500
1	35000	13425.1	500

- Average Annual Income is highest for Interest Rate category 1 and there is no difference between Interest Rate Category 2 and Interest Rate category 3

Average Annual Income is highest for Interest Rate Category 1 and almost same for Interest Rate Category 2 and Interest Rate Category 3

Interest_Rate	Maximum	Average	Minimum
1	4900000	82278.76	5000
3	7500000	72551.85	6400
2	7141778	72368.78	4000

- There is Multi collinearity in dataset and there is high correlation i.e. almost correlation coefficient equals to 1 that means a strong positively linear relationship between two features Open Accounts and total accounts. To overcome this, we can create a new feature in our dataset either by adding Open Accounts and Total Accounts or we can use Open accounts itself because those were currently active accounts

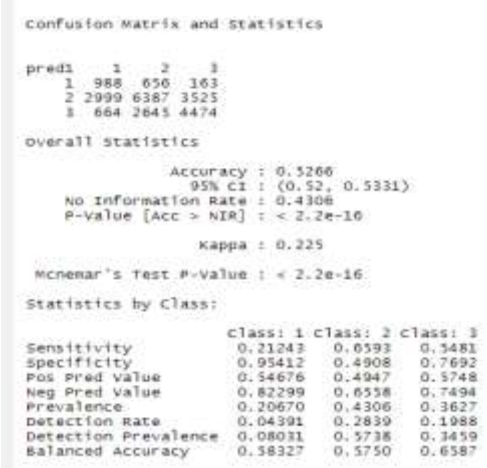


Empirical Analysis:

I’ve used Basic Multinomial Logistic Regression, Random Forest, Linear Discriminant Analysis to predict the Interest Rate Category and feature selection techniques such as forward selection and backward selection using regsubsets from leaps package on Multinomial Logistic Regression. The results are tabulated below for Basic Multinomial Logistic Regression

Multinomial Logistic Regression:

Accuracy	0.5266		
No Information Rate	0.4306		
Metrics	Classes		
	Class 1	Class 2	Class 3
Sensitivity	0.21243	0.6593	0.5481
Specificity	0.95412	0.4908	0.7692
Balanced Accuracy	0.58327	0.575	0.6587

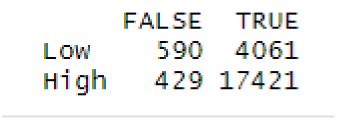


```
Confusion Matrix and Statistics


pred1    1    2    3
    1  988  656  163
    2 2999 6387 3525
    3  664 2645 4474

Overall Statistics

               Accuracy : 0.5266
                 95% CI : (0.52, 0.5331)
    No Information Rate : 0.4306
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.225

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity           0.21243   0.6593   0.5481
Specificity           0.95412   0.4908   0.7692
Pos Pred Value        0.54676   0.4947   0.5748
Neg Pred Value        0.82299   0.6558   0.7494
Prevalence            0.20670   0.4306   0.3627
Detection Rate        0.04391   0.2839   0.1988
Detection Prevalence  0.08031   0.5738   0.3459
Balanced Accuracy     0.58327   0.5750   0.6587
```

As I am primarily interested in identifying the Low Interest Rate Category applicants and High Interest Rate Category Applicants correctly, I’ve used One Versus All Approach i.e. considering all other classes i.e. Medium and High as High Classes in order to correctly identify the Low Interest Rate Category applicants and adjusted threshold in Logistic Regression Model to correctly identify the Low Interest Rate Category. The confusion Matrix is tabulated below and we used Sensitivity as metric for identifying Low Interest Rate Category observations correctly

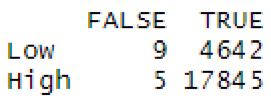
One Versus Rest – Low Interest Rate Category:

Default Threshold 0.5



```
      FALSE  TRUE
Low     590  4061
High    429 17421
```

Threshold 0.2



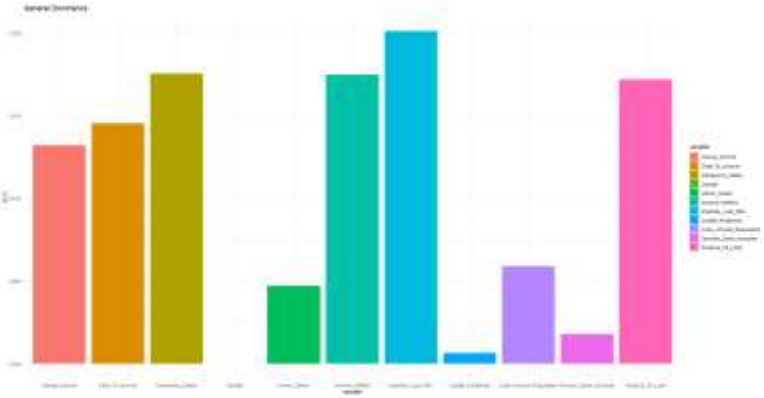
```
      FALSE  TRUE
Low       9  4642
High      5 17845
```

Sensitivity = 4061/4061+590 = 0.8731

Sensitivity = 4642/4642+9 = 99.80

We used Sensitivity as metric for identifying observations or loan applicants which belongs to category Low Interest Rate and at threshold of 0.2 we are able to get 99.8% Sensitivity i.e. Out of actual Loan Applicants having Interest Rate Category Low we are almost identifying or classifying everyone correctly.

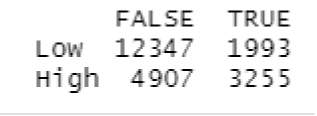
Feature Importance:



One Versus Rest – High Interest Rate Category:

For identifying all High Interest Rate Category applicants accurately, I’ve used One vs All approach by considering Low and Medium as one Class and High as one Class and adjusted threshold to accurately identify all High Interest Rate Category applicants.

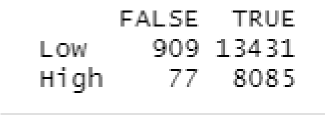
Default Threshold 0.5:



```
      FALSE  TRUE
 Low  12347  1993
 High  4907  3255
```

Sensitivity = 3255/8162 = 0.3987

Threshold 0.1:

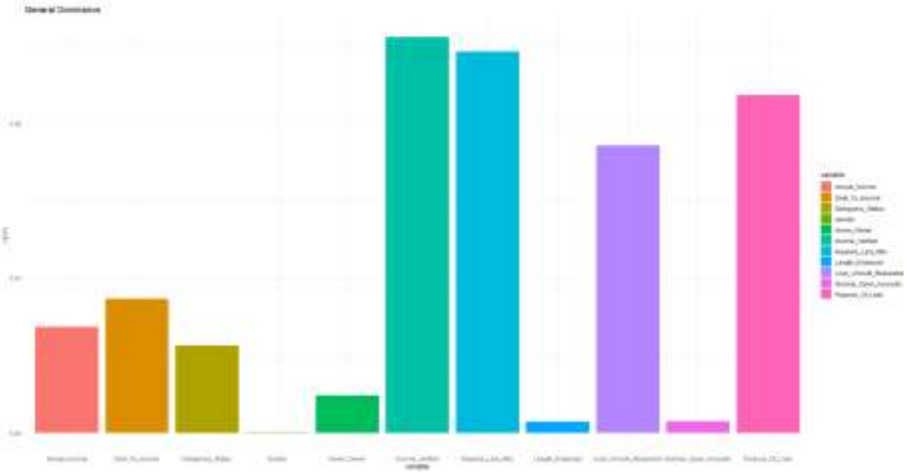


```
      FALSE  TRUE
 Low    909 13431
 High    77  8085
```

Sensitivity = 8085/8162=99.05

I’ve used Sensitivity as metric for identifying observations or loan applicants which belongs to category High Interest Rate and at threshold of 0.1 we are able to get 99.05% Sensitivity i.e. Out of actual loan applicants having High Interest Rate Category we are almost identifying or classifying everyone correctly.

Feature Importance:



Conclusions:

Using Loan Application dataset from Analytics Vidya, we analyzed the applicant's preferences and how different factors are affecting Interest Rate Category and are able to correctly identify or Classify applicants belonging to Low Interest Rate Category and High Interest Rate Category with Sensitivity of 99.08% and 99.05%. Also, from dominance analysis plots Delinquency Status is having significant influence in deciding Interest Rate Category Low and Gender is not having any impact in deciding the Interest Rate Category Low but having a little impact in deciding the Interest Rate Category High

Source:

JanataHack Machine Learning for Banking "datahack.analyticsvidhya.com/contest/janatahack-machine-learning-for-banking/"