LOAN CREDIT-WORTHINESS CLASSIFICATION

-Prajwal Waykos

- +917249542810

Pwaykos1@gmail.com

Project Repo = https://github.com/Praj-17/Loan-Creaditworthiness-classification

Colab = https://colab.research.google.com/drive/1NiH_xen-GZwkCAk34TbK8NpZB-2-HJiC?usp=sharing

Drive = https://drive.google.com/drive/folders/1-U5q9mKspe22HN_4fvZhVFGZ0zpKI7AS?usp=share_link

Problem Statement

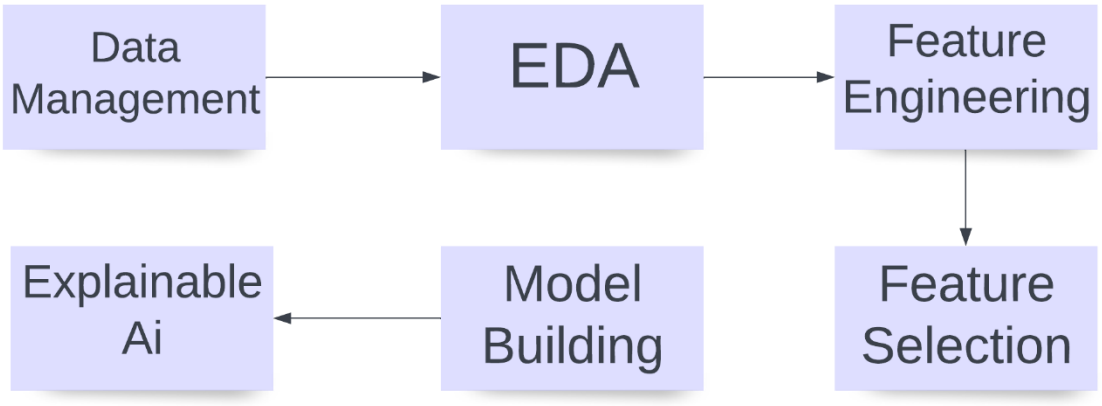
The Problem statement is to prepare a model based upon the given dataset which classifies a new applicant as risky or creditworthy.

Abstract

The dataset presents us with 2 tables namely applicant and loans which contain applicant and loan data respectively. The Dataset is available in the repository.

I have performed all the major Machine Learning procedures such as , Exploratory Data Analysis, Data Analytics, Data Preprocessing, Feature Engineering, Feature Selection, Modeling and Explainable AI on the dataset.

Procedure



Methodology

1. Data Management – visit :

This file is used to merge the given two tables and also to generate, Pandas Profiling report everytime needed.

Components:

1. Merging of 2 tables
2. Generating reports of the dataset.

1. Merging of 2 tables:

- There are 2 tables given, Applicant and Loan which are separated logically by the data they contain.
- But for EDA and Better understanding of the model we will be joining the 2 tables based upon the relationships in them.
- In the Applicant Table we can see the first column as applicant_id which is the **Primary key** of it.
- Moreover, in the Loan Table we can see the Second column as applicant_id which is the **Foreign Key** of it.
- Based upon this common column we will be joining the 2 datasets and store in a csv file called data_merged.csv.

2. Generating Reports of the dataset.

Pandas profiling report is library is an automated python tool for basic EDA.

- This report provides basic EDA and an overview of the dataset.
- The following important data such as

    - Missing Data
    - Types of Variables
    - Classes in Ordinal variables
    - Range of Numeric Variables
    - Correlations and Sample Data Visualizations

You can visit all the reports created here.

Some Inferences from the generated report

- There are a total of 27 columns and 1000 records of each column.
- Luckily there are no duplicate reocords or columns in the dataset.
- There are 4 numeric and 23 categorical features in the dataset as of now
- he Balance of the dataset is 7:3 which makes it critical to think weather it should be balanced or imbalanced. Since we lack enough data, we would be considering it as imbalanced and performing oversampling in the later phases.
- Around 12% of the data is missing.
- There are some columns which have more than 60% percent of missing cells. We might need to eliminate them or they will add unnecessary bias in case we try imputing them
- All sorts of Features are present within the dataset and hence, we need to take special care while handling each variable
- The Variable "Telephone" is constant (No Use)

2. Feature Engineering:

EDA or Exploratory data analysis is a procedure in Machine Learning where we analyze the given data both statistically and logically to generate insights that could solve difficult business problems.

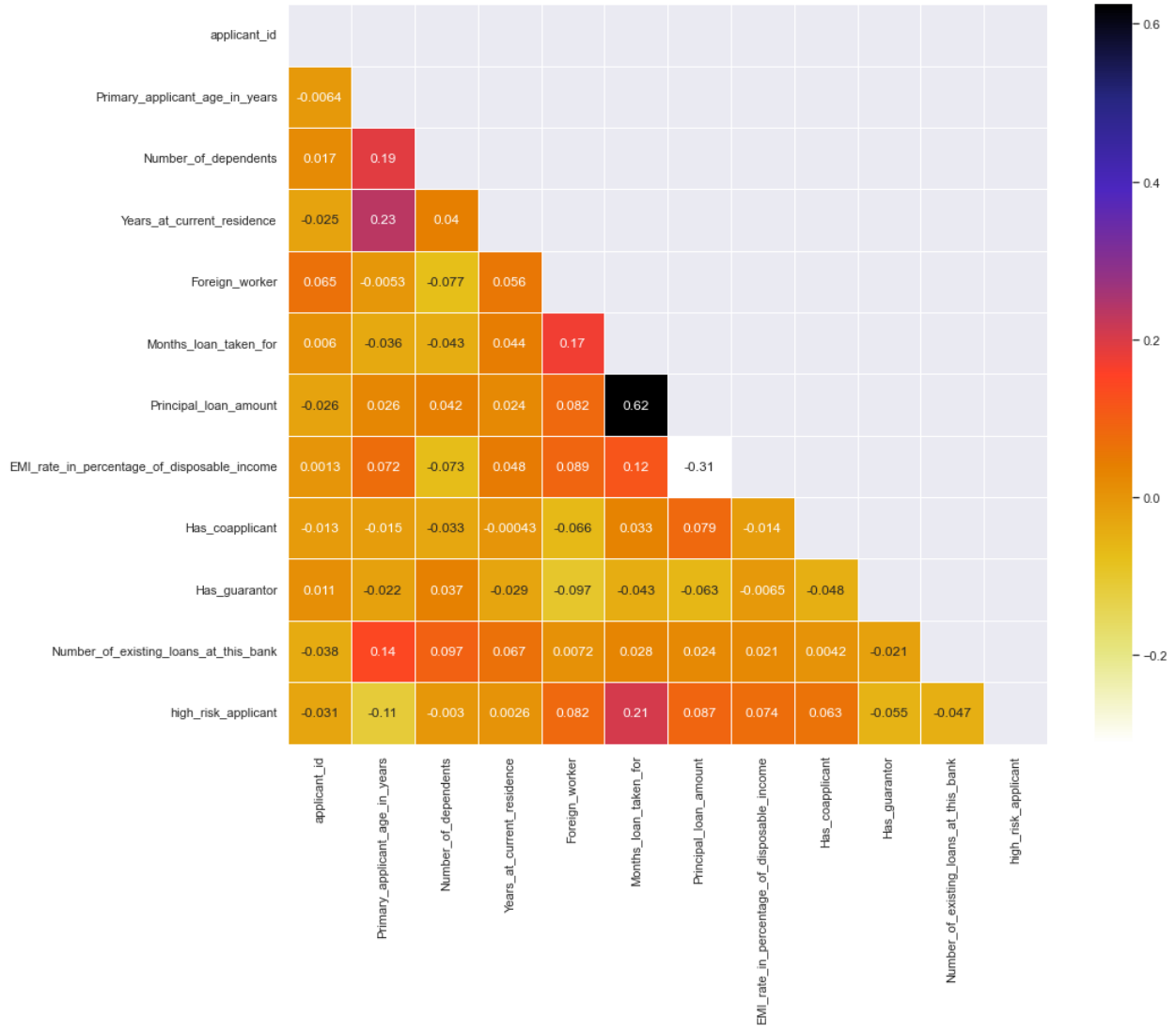
I have performed EDA in 2 phases

1. EDA and FE notebook –
2. Advance EDA notebook –

1. EDA and FE -

- Since the basic , eda has been covered by the generated report we will see some visualizations. To better understand the spread and variance of data.

Correlations



The chart above states various correlations among the variables. But there is one thing to note here that it consists only 13 cols and not the remaining columns. This is because the other cols might be having some string value or datetime values within them. Hence, We should consider Data Cleaning/ and Feature Engineering before thoroughly understanding the data.

Insights of the correlations from raw data.

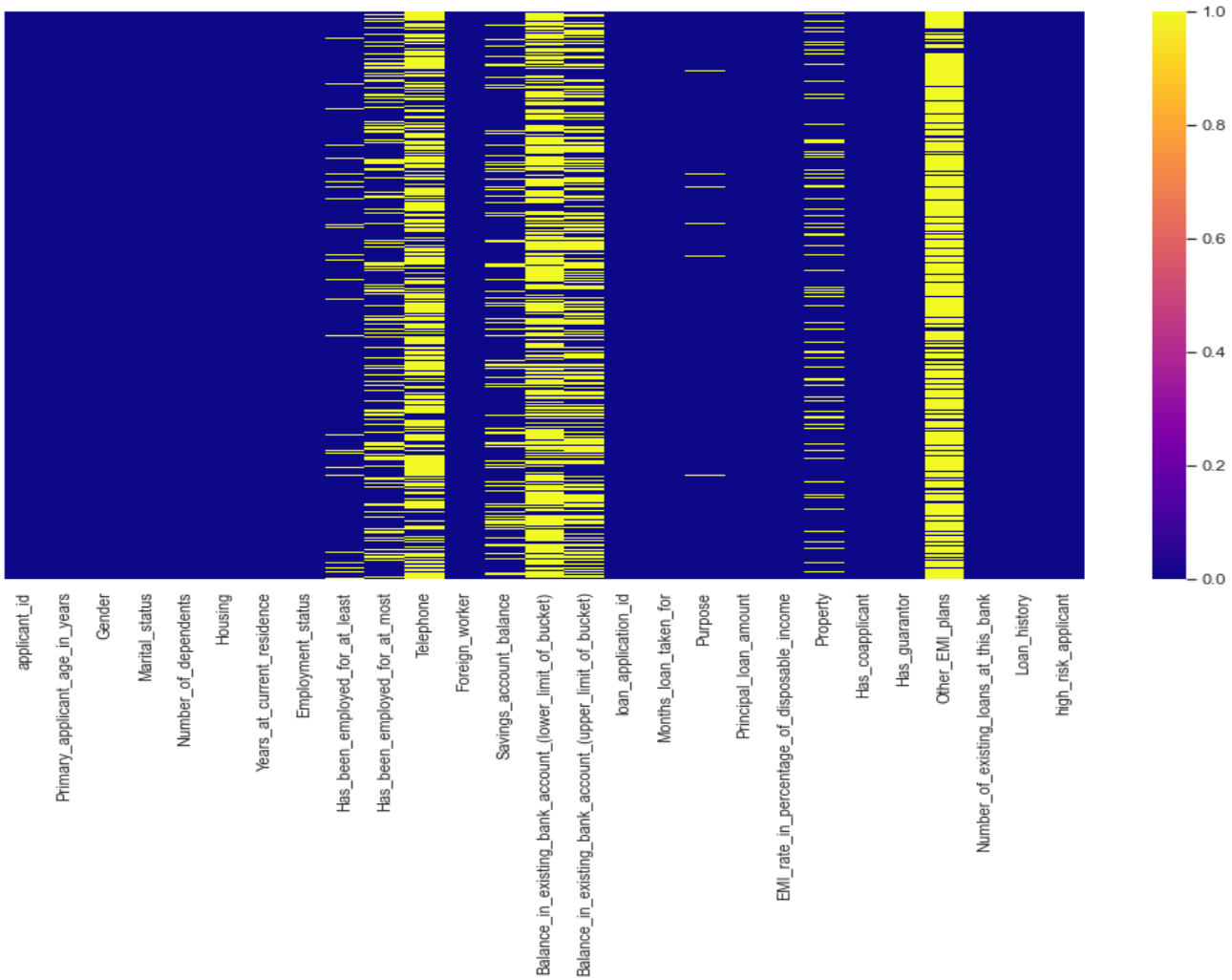
- Duration for which the loan was taken is highly correlated with applicant being a defaulter.
- The more is the age of the applicant the more loans he might take
- The more is the principal amount the Longer it gets.
- The more is the age of the applicant the longer he is supposed to live at the same residence.
- Foreign Workers usually take bigger and longer loans.
- The More the dependents the more loans a person takes
- The more the applicants age the more people depend on him.
- We must also observe that there are 12 columns in the above plot which means there are 15 columns with no-numeric datatype.

All the above made conclusions are logically correct and hence, we can also conclude that the data is natural and not artificially made.

Feature Engineering

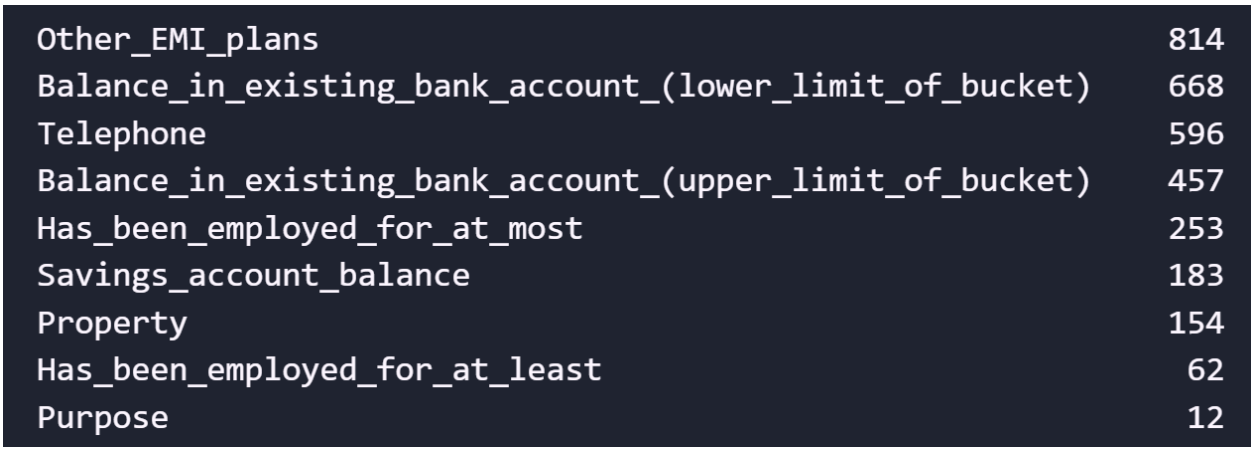
1. Missing Data –

The following plot gives us insights on which columns have missing cells and with what quantity they are missing.



From the above Data we can conclude that

- There is total 9 columns with null values
- Out of the 9 columns 2 Columns namely 'Telephone' and 'Other_EMI_plans' needs to be eliminated entirely. Since, Telephone is a constant and Other_EMI_Plans have more than 80% null Values
- The 2 columns 'Existing Bank Balance' also needs to be eliminated and they also do not enough variance among the data.



Other_EMI_plans	814
Balance_in_existing_bank_account_(lower_limit_of_bucket)	668
Telephone	596
Balance_in_existing_bank_account_(upper_limit_of_bucket)	457
Has_been_employed_for_at_most	253
Savings_account_balance	183
Property	154
Has_been_employed_for_at_least	62
Purpose	12

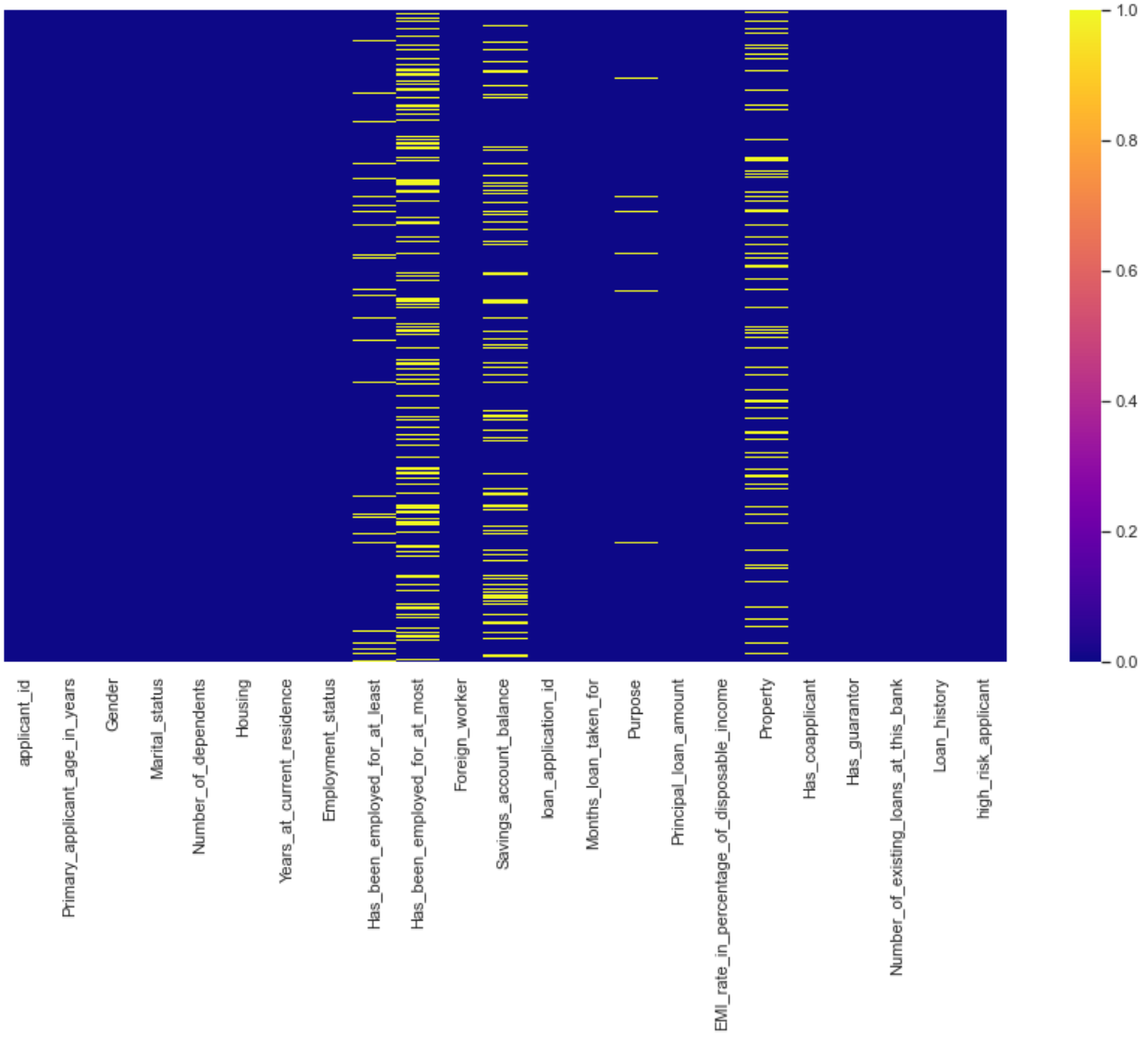
The Picture above enlists the columns having null values in descending order.

As you can see there are 9 features with missing values most of which are non-Numeric.

Hence we would need to apply Imputation where we have at least 60% of the data available and drop the columns with more than 40% of the missing Data.

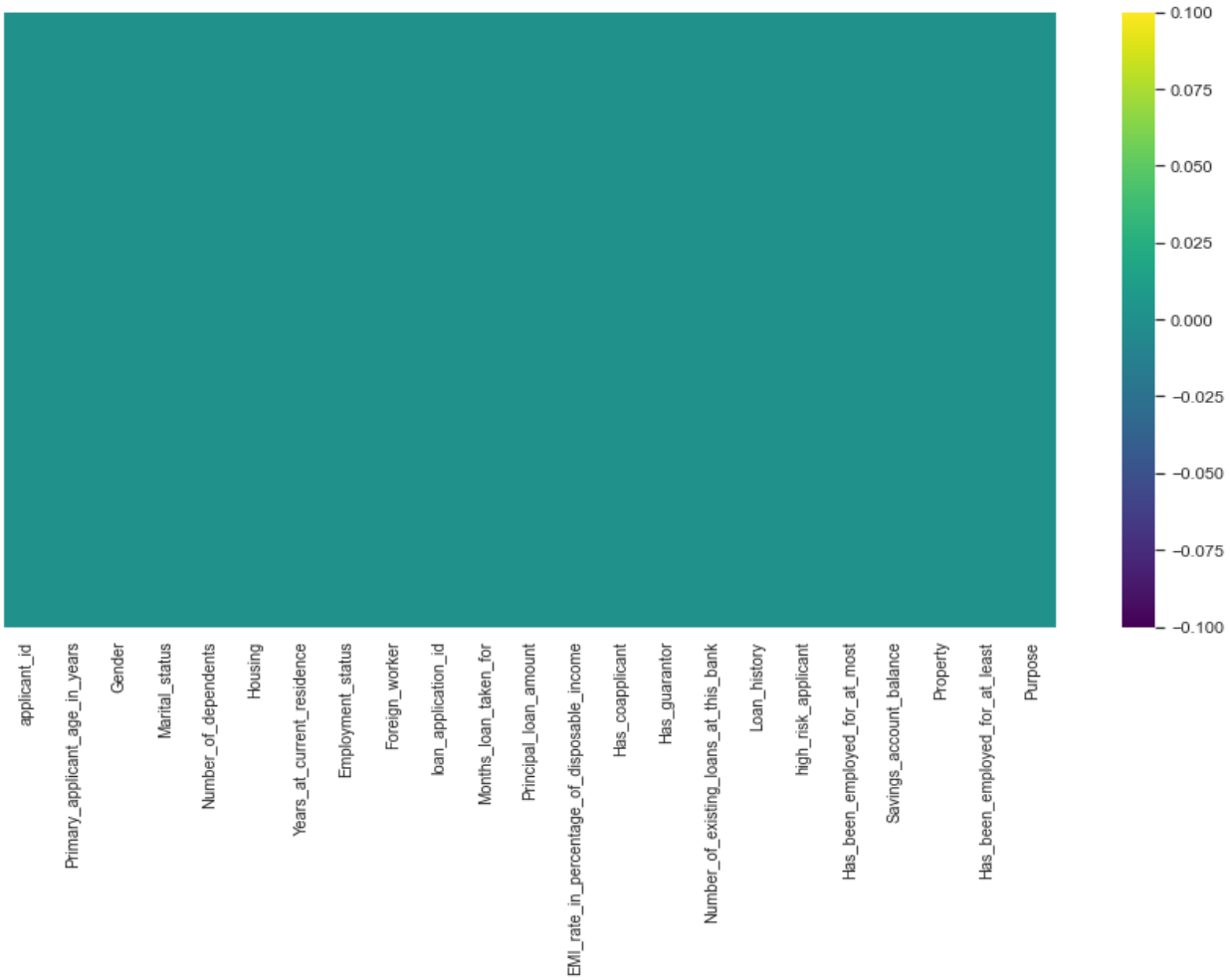
As you can see there are 9 features with missing values most of which are non-Numeric.

Hence we would need to apply Imputation where we have at least 60% of the data available and drop the columns with more than 40% of the missing Data.



After Eliminating top 4 columns with missing values

The Picture above depicts that after eliminating the top 4 columns the density of missing values has certainly decreased.



The chart above represents that there are no null values in the dataset.

Data Imputation –

I have used SK Learns KNN imputer class for imputing the missing cells.

The value of k is 5 which mean is will take the weighted average of 5 nearest neighbors as the missing value and also, fill it in the missing cells.

2. Handling Categorical Features

The dataset consists of 8 categorical features, as listed below.

The categorical features are further classified as

a. **Ordinal** - Ordinal features are the features where order matters in the categories
b. **Nominal** – Nominal features are the features where order does not matter in the categories.

```
Nominals = ['Gender', 'Marital_status', 'Housing']
Ordinals = ['Employment_status', 'Savings_account_balance',
'Purpose', 'Loan_history', 'Property']
```

According to my research on the Problem Only the Above Features are Nominals or Ordinals and the remaining ones are Numeric, just the type is object.

Now, We will be Applying One Hot encoding to the nominal features and Factorizing the Ordinals variables according to their rank.

Find the table of the associated ranks of each category with respect to the classes here.

Reamaining Categorical Columns

Now there are only 3 Categorical Columns Remaining

- Application Id
- Least Employement
- Max Employement

Out of these Application Id would be removed since it is a primary key. The Remaining 2 columns could be merged into one column as Average Employment by taking the average of the 2

Hence, Here we come to our first derived column called Average Employment

Average Employment = Max employment + Min Employement /2

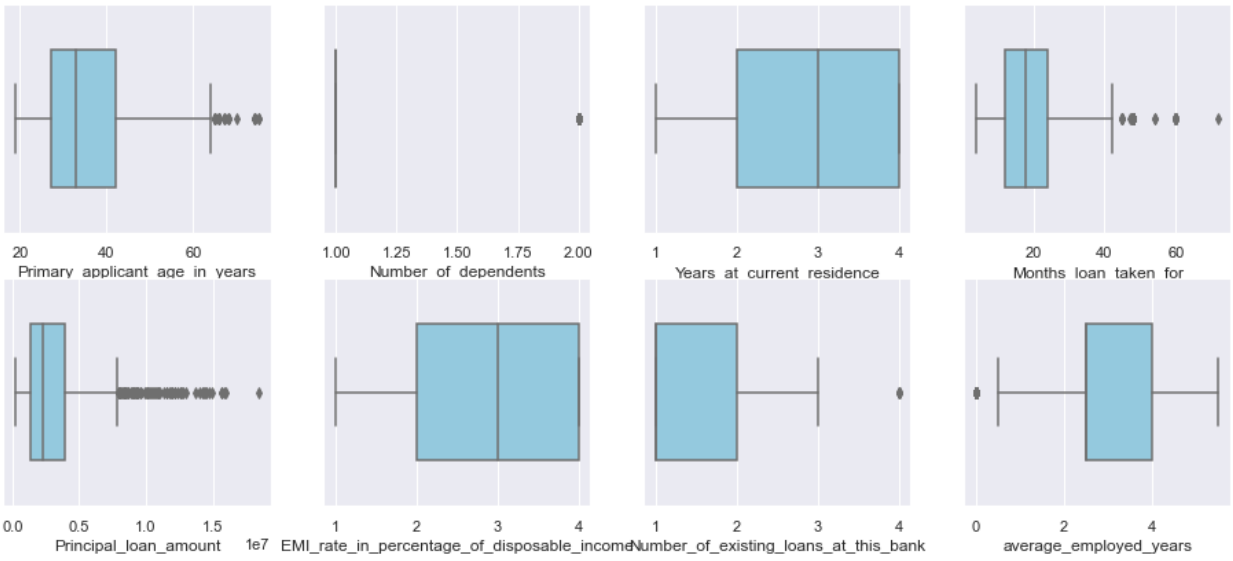
3. Looking out for outliers:

- Outliers are any arbitrary values which are at extreme ends of the range of the feature.

- I have used a statistical method called Z-Score and standard deviation to detect outliers In the dataset.

- If a value is farther from the 3rd standard deviation of the standard normal distribution of the data, it is considered as an outlier.

- The Following Boxplot provides better insights on the spread of the data.



Boxplots representing the outliers in the features.

Observations

- From the above plot we can infer that there are very few features having significant outliers
- The Features Principal , Age, Months loan taken far have some significant amount of Outliers
- Since, we would be Using Randome Forest Classifiers and XG Boost Models the outliers might not affect significantly

Imputation of outliers

```
[60, 60, 60, 60, 60, 60, 60, 60, 60, 72, 60, 60, 60, 60]
[70, 74, 75, 74, 75, 74, 74]
[12579000, 14421000, 12612000, 15945000, 11938000, 14555000, 12169000, 11998000,
13756000, 14782000, 14318000, 12976000, 11760000, 12389000, 12204000, 15653000,
14027000, 14179000, 12680000, 15857000, 11816000, 15672000, 18424000, 14896000,
12749000]
```

The picture shows the outliers found.

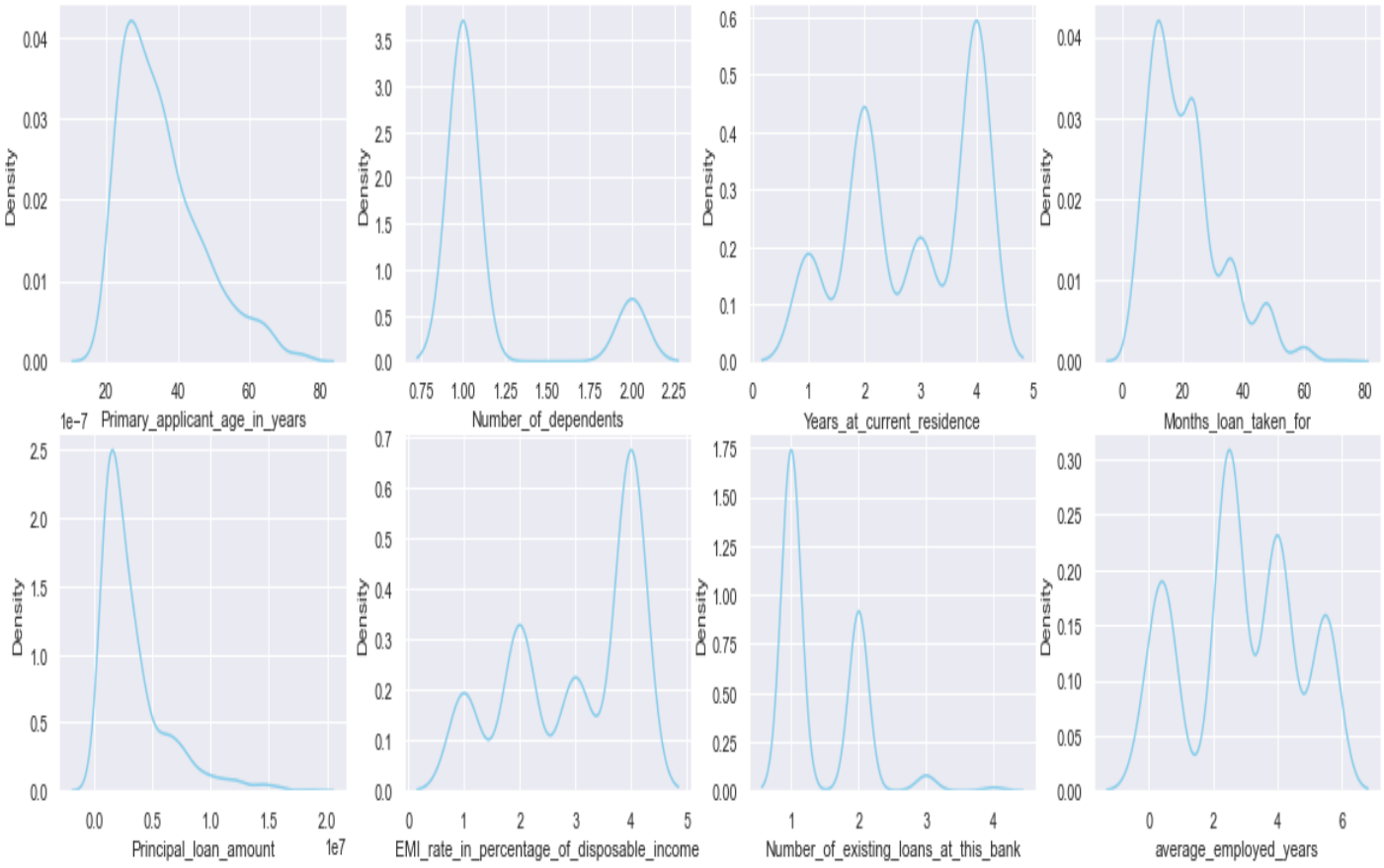
Hence, we found that there are few values in 2 features which require some imputation but since the features are correlated and changing/imputing values would have a significant effect we would not impute anything and leave the values as it is.

Moreover we will be using Logistics Regression and Random Forest classifiers, both of which are not affected by Outliers and hence having few outliers won't create any problem

4. Feature Transformations

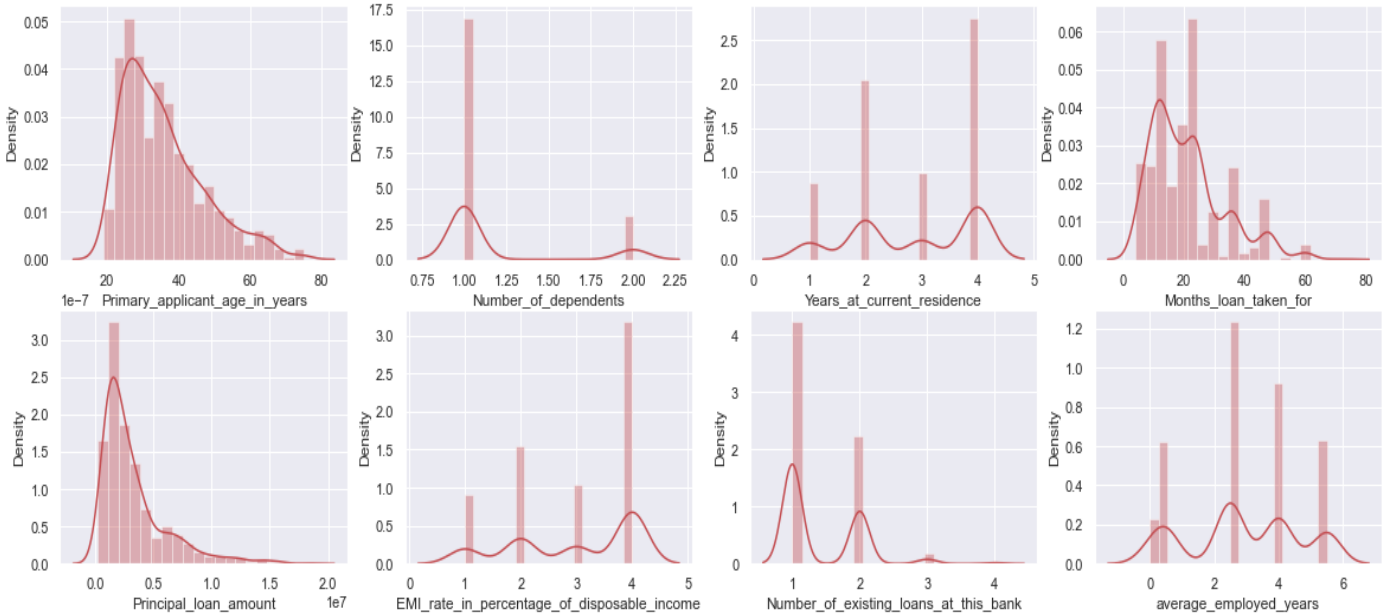
In this module we will be looking out for various distributions the numeric features are having.

The following is the KDE plot of various numeric features in the dataset.



Distribution of numeric features.

The following chart gives even better view of the distributions.



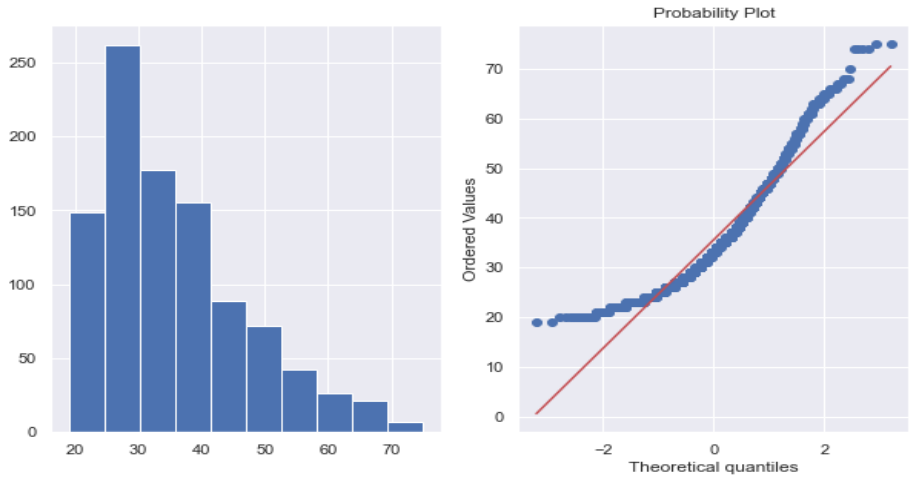
Observations from above Data

- The columns age, months and principal do seem to have normal distribution and the other columns are sufficiently discrete.

- Although age, months and principal are normally distributed, they are all **left skewed** and hence, we will be transforming them to form a proper **Normal Distribution.**

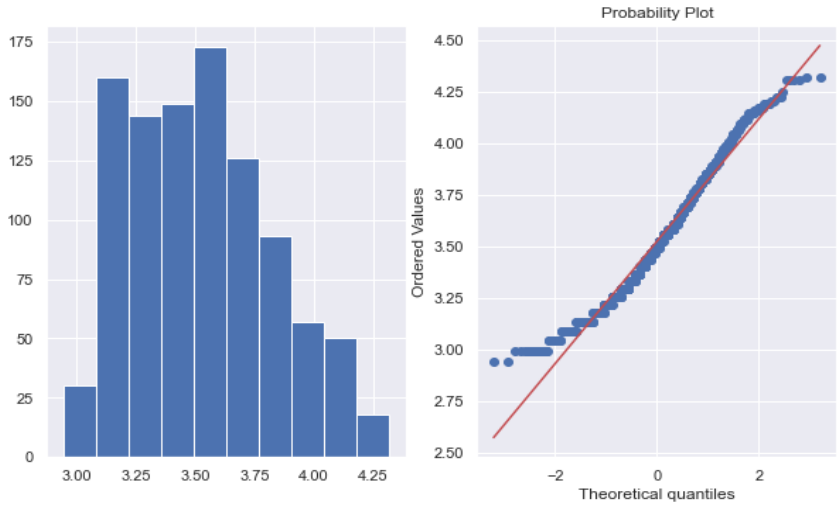
In statistics we have a method called, QQ plot which visualizes the data on a straight line,

Following chart shows the QQ plot applied on one of the columns.



QQ plot of age of the applicants

Since, the column is left skewed we will apply log transformation on the dataset and convert it to normal form



QQ plot of transformed data

As you can see taking the log of values that is converting them to lagrithmic distribution makes the data even more normally distrubuted. Hence, we will be applying the same method to other 2 features as well

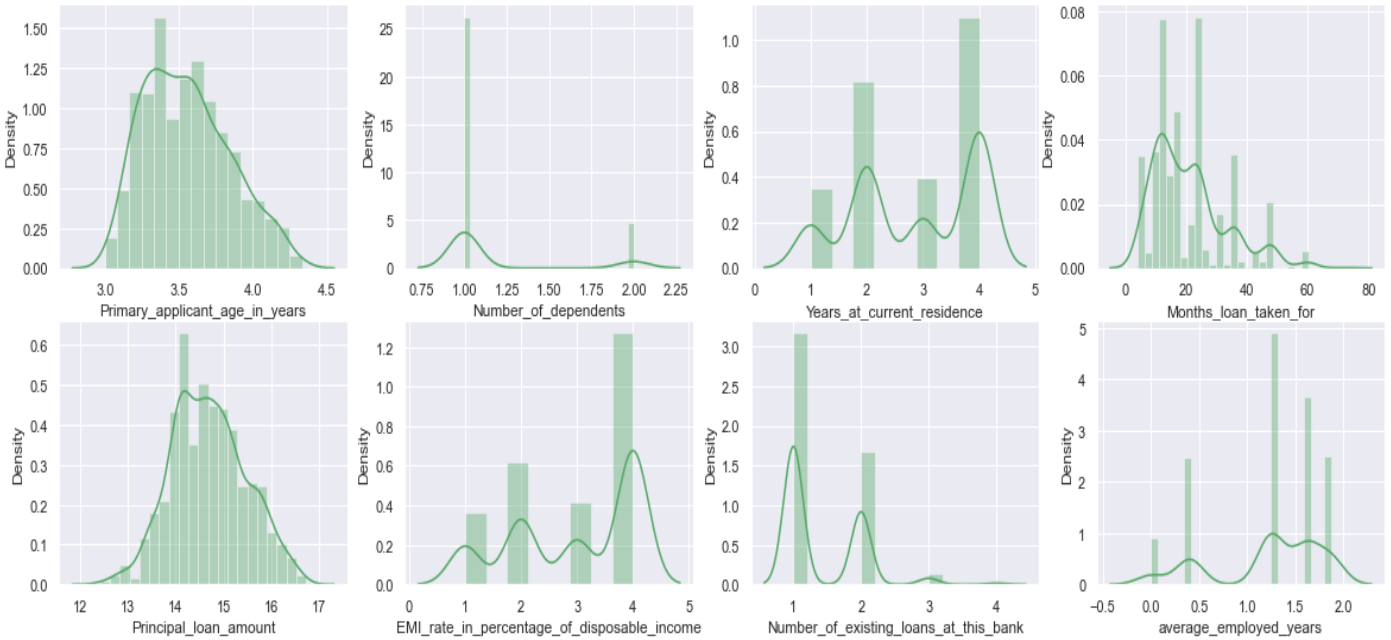


Chart after transforming the 3 left skewed variables

The difference in 3 transformed features is loud and clear.

The Complete Transformed Data is Stored here.

5.Advance EDA

In this notebook we will be Drawing out various insights, plot useful visualizations and also derieve some new features if needed

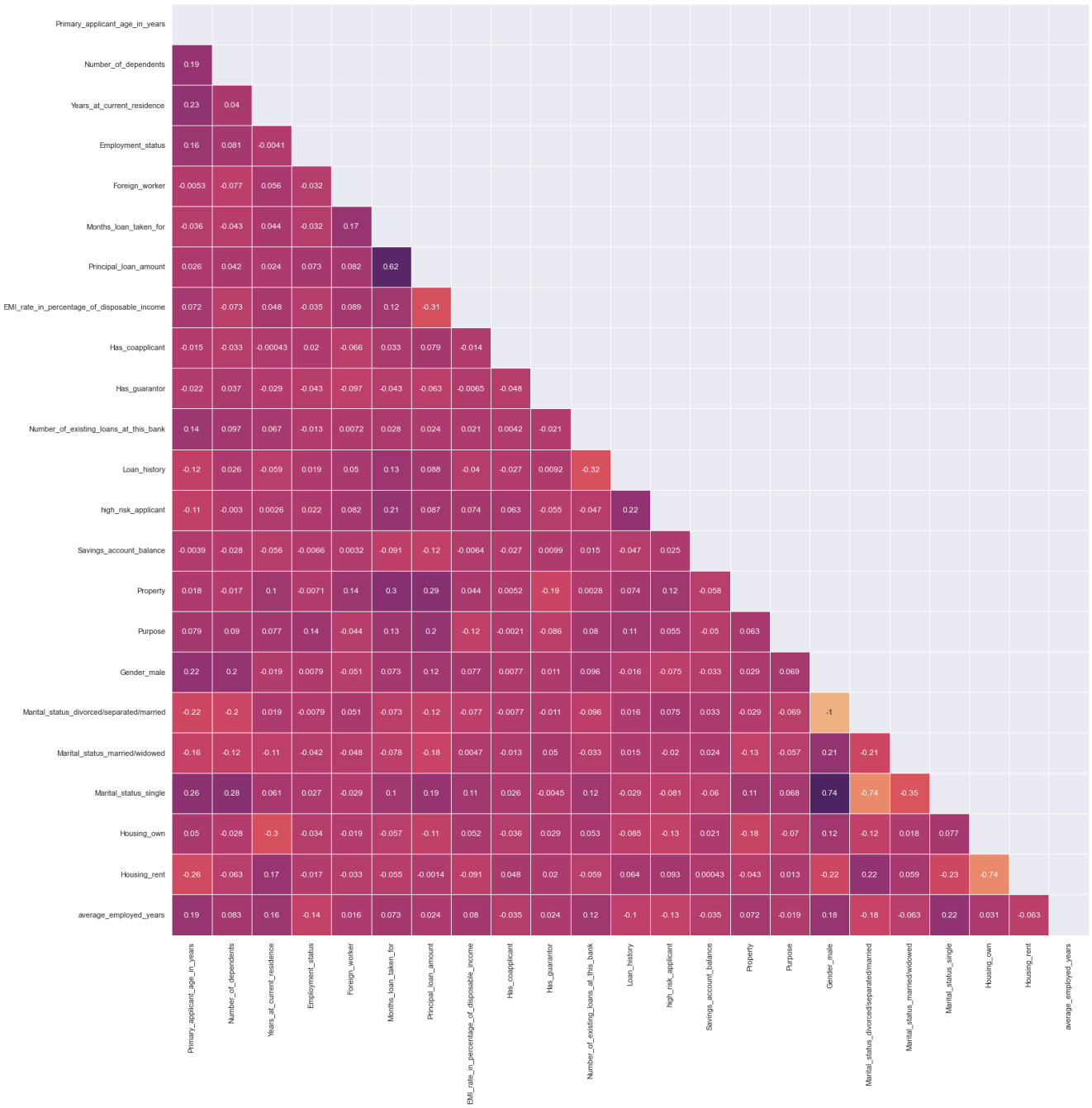
Bivariate Analysis

Correlations

- Why spearman correlation and why not Pearson?
- Because Person only considers linear relationship among the variables that is it looks only for a constant increase/decrease with respect to the other variable.
- Whereas spearman also, takes care of monotonic relationships

Some Conclusions Drawn

- Risk is directly affected by the applicants History.
- Males have more people dependent on them
- There are more single males applying for loan.



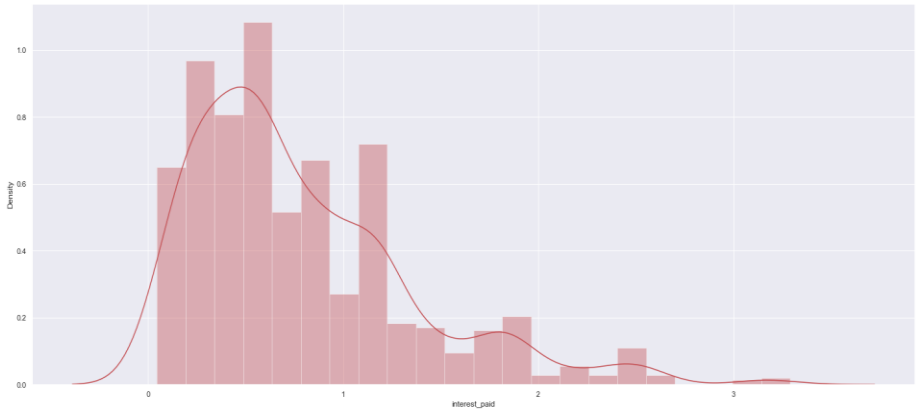
Refer GitHub for proper view

Now we will be considering a lot of new features and see how they affect the target feature

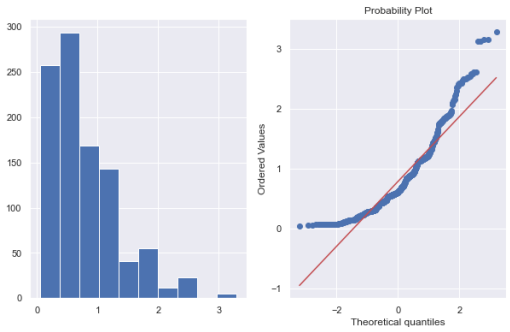
1. Interest Paid

This is the second calculated column that I have calculated, the Interest paid, is calculated by the following formula.

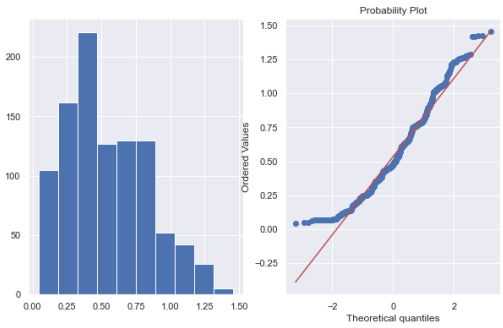
Interest paid = (Principal * number of years * Rate)/100



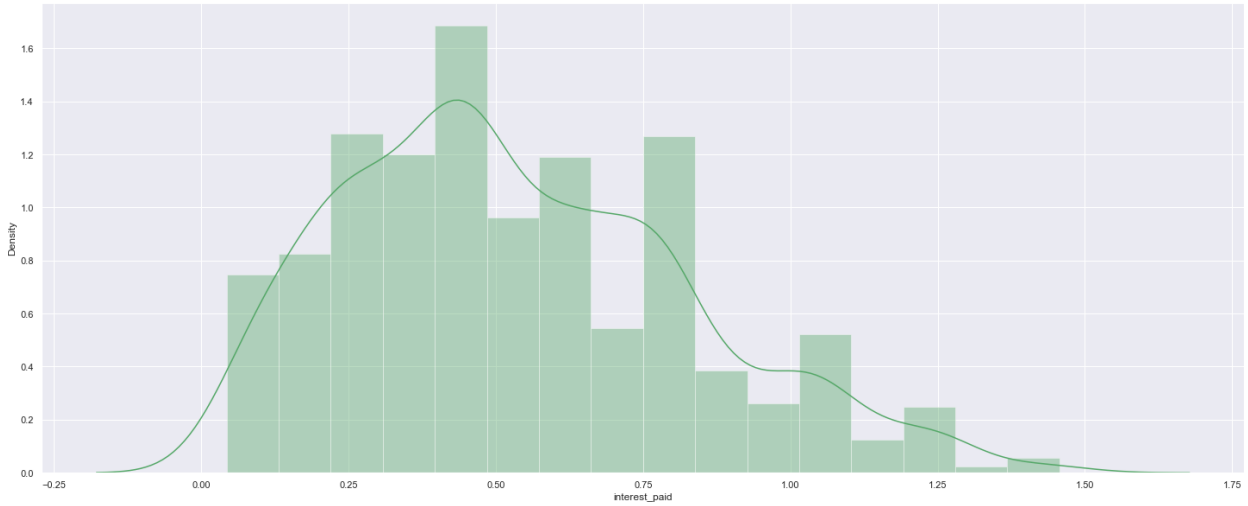
Distribution of the interest column



QQ plot of interest



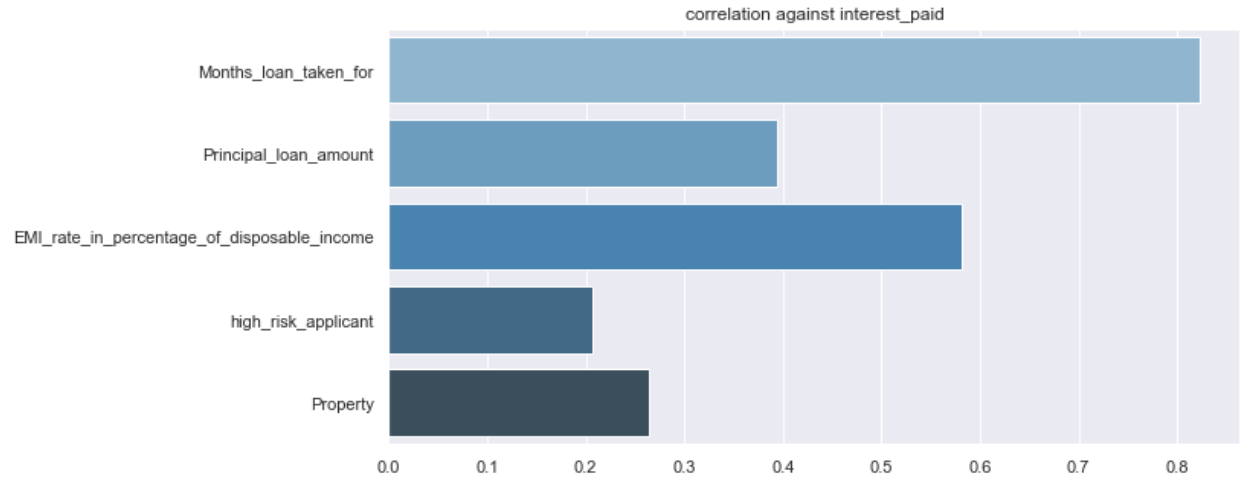
QQ plot of transformed interest



The distplot for the transformed variable, now it looks more normal than before.

Now it seems comparatively more normal, we can bring it to exact normal distribution by a series of log transformation but it would be prone to big statistical blunders and hence we would be keeping it like this.

Correlation of interest paid with other features.



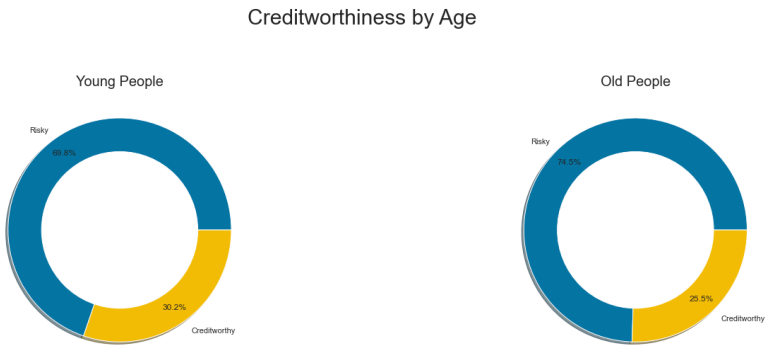
Plot measures the correlation of interest paid with all other features and sows the significant ones

As we can see the Parameters used to construct the feature are significnatly related to the interest and hence, these 3 columns can be easily removed.

We will segment the applicants based upon various features and check which type of people are more creditworthy.

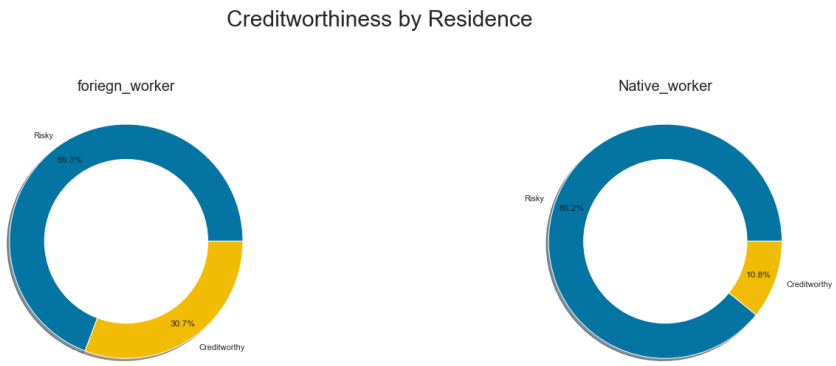
Note – Remember the dataset is imbalanced towards the risky category and hence, even a slight difference of 1-2% is significant.

1. Based Upon Age.



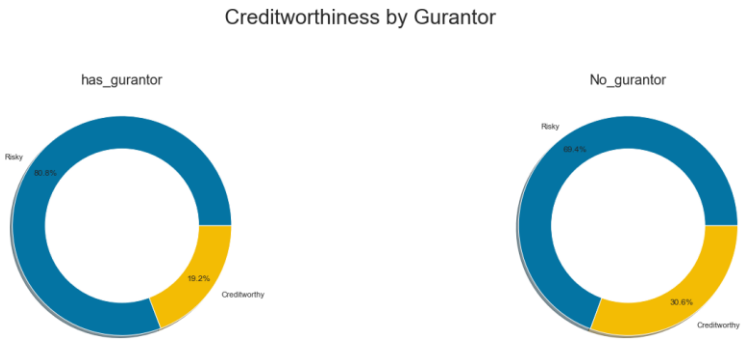
From the Above Charts We can Conclude that Young peple are more creditworthy

2. Based upon



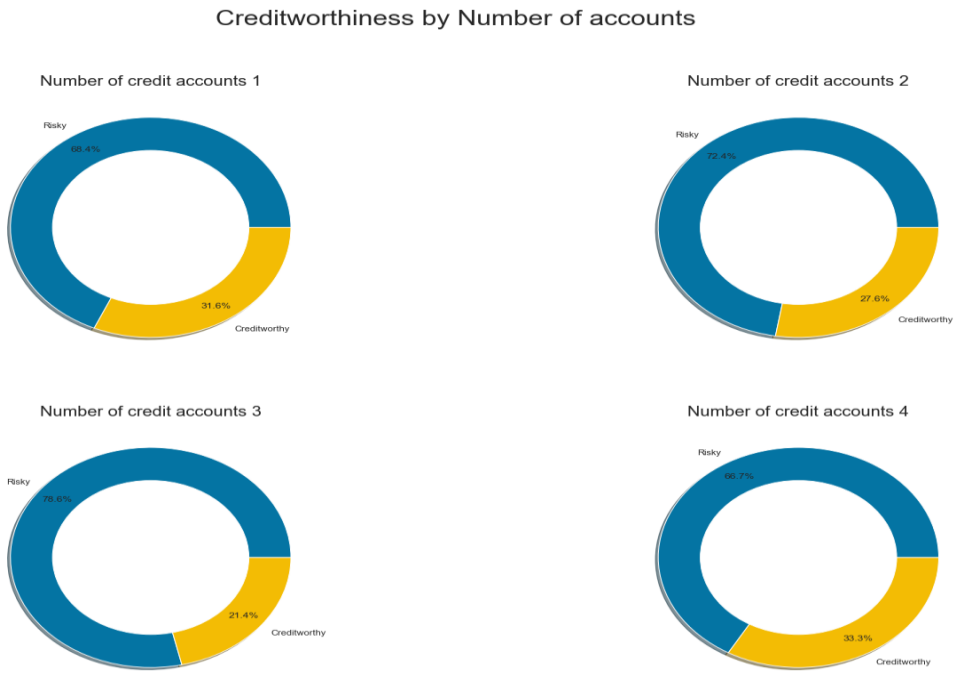
From the Above Charts We can Conclude that Foreign Workers are more creditworthy

3. Based upon Gurantor.



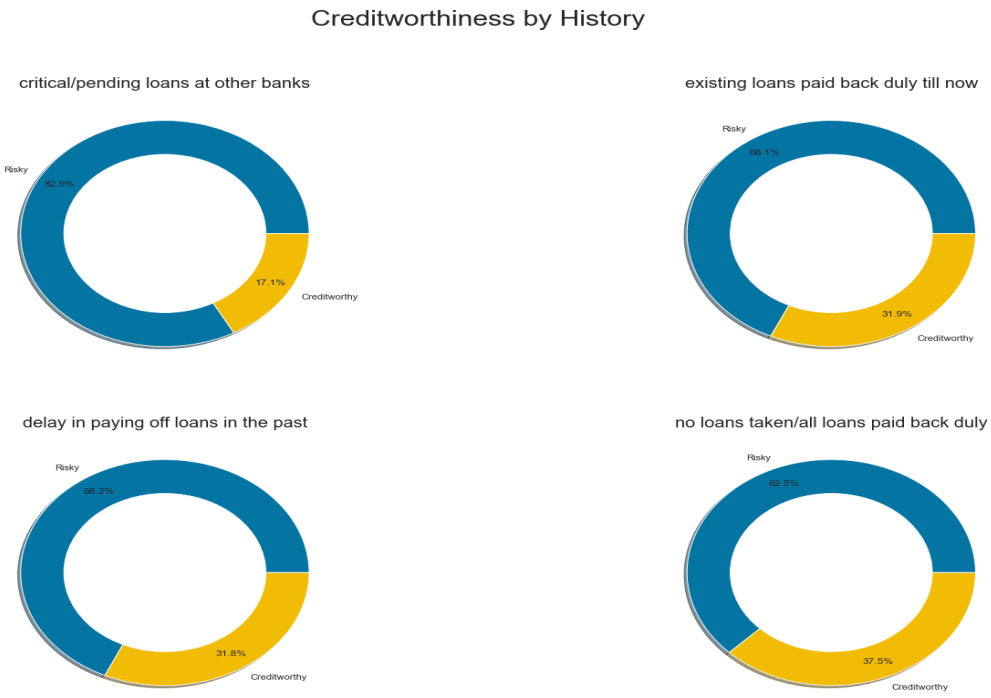
From the Above Charts We can Conclude that people **not** having Guarantors are more creditworthy.

4. Based upon Number of credit accounts.



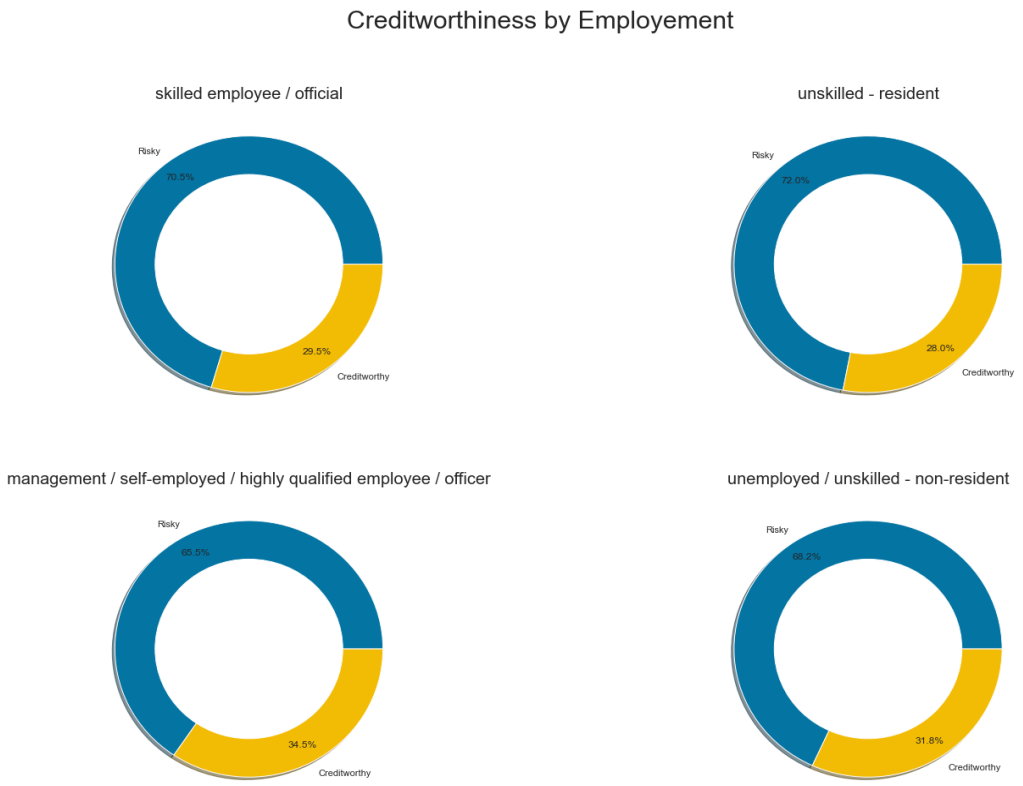
From the above charts we can conclude that people with lesser number of credit accounts are more creditworthy.

5. Based Upon Credit History.



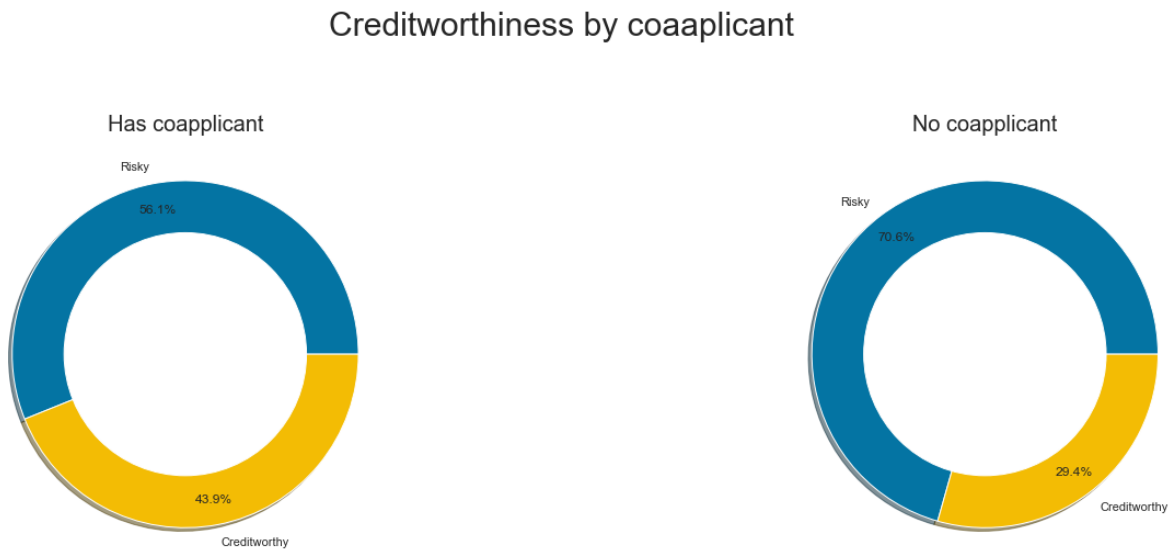
From the above charts we can conclude that the loan history is also a significant feature and people who have pending loans with other banks must be avoided.

6. Based upon employment:



From the above chart we can conclude that Unskilled employees should be avoided since they are less creditworthy.

7. Based Upon Co-applicant.



Hence, from above chart we can conclue that having a copplicnat is more creditoworthy

8. Based upon Purpose

Education loans have the least probability of being default followed by new vehicle and loans taken for used vehicles, and electronics vehicle are at higher risks

Conclusions

- Education loans have the least probability of being default followed by new vehicle and loans taken for used vehicles, and electronics vehicle are at higher risks
- Unskilled employees should be avoided since they are less creditworthy
- The loan history is also a significant feature and people who have pending loans with other banks must be avoided.
- As you can see High risk is related to interest paid and longer loan duration. Hence, we can conclude that giving loan with higher interests or for longer duration increases the risk.
- Young people tend to take more loans and are more creditworthy than old people.
- Having a co-applicant is more creditworthy
- Foreign Workers are more creditworthy as compared to native workers.
- Surprisingly, People **not** having guarantor are more also more creditworthy.
- People with lesser number of credit accounts are more creditworthy.

Whose loan should be approved?

A skilled young person taking an education loan who lives in foreign and has a co-applicant with fewer/no number of credit accounts is the best candidate to provide a loan.

Followed by the permutations of the above conclusions.

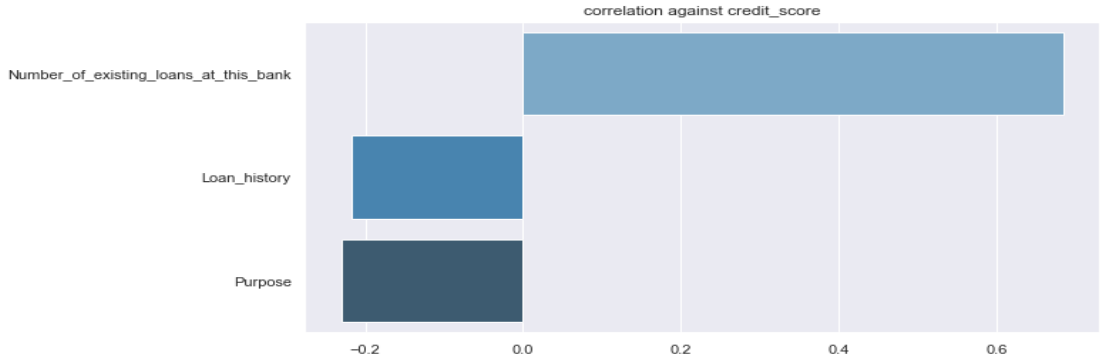
Credit score derivation

From the above conclusions we can derive something like credit score, which we consider all above points. The credit score will be responsible parameter for deciding the approval of the application

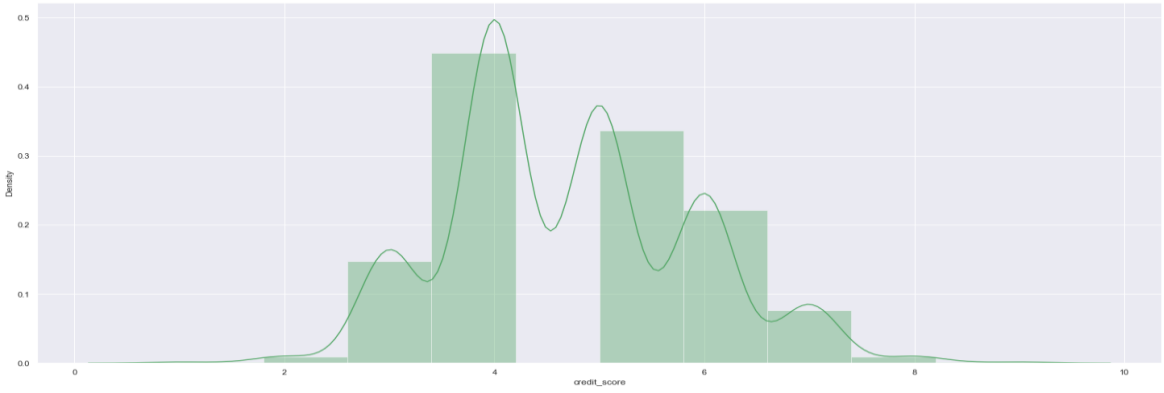
Credit score = [Loan Purpose + skilled + good history + comparatively shorter duration + young + foreign worker - gurantor + no. of credit accounts + has_copplicant]

- if loan purpose is in top 4 most creditworthy purposes, we will add one to credit score
- if the applicant is skilled, we will add one point
- Loan history gets 1 point
- similarly, we add other parameters
- since guarantor is inversely proportional, we will subtract it.

Correlations of credit score



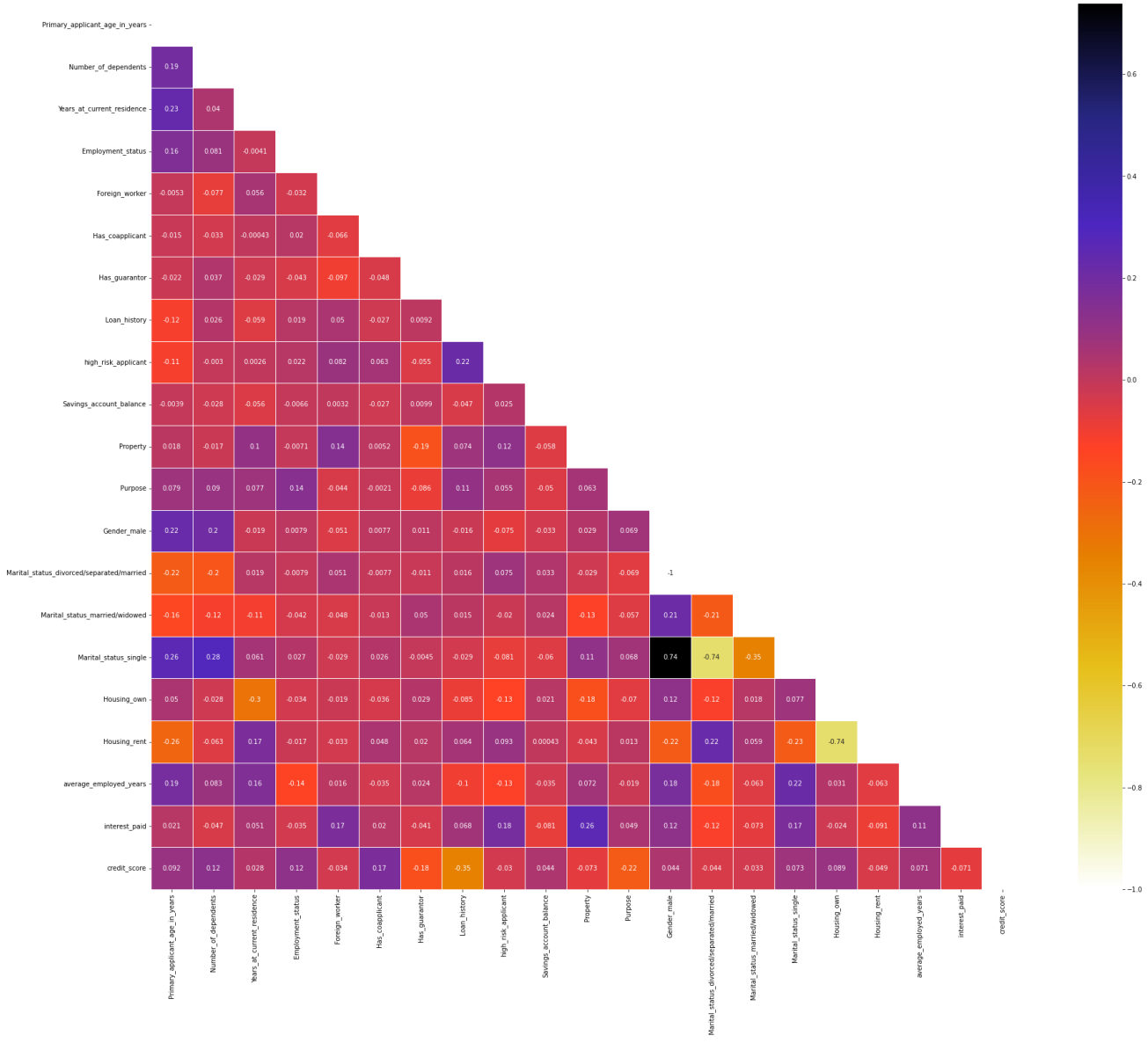
Most significant correlations of credit score.



Credit score is already normally distributed

Feature Selection

Looking for correlations in the Clean data.

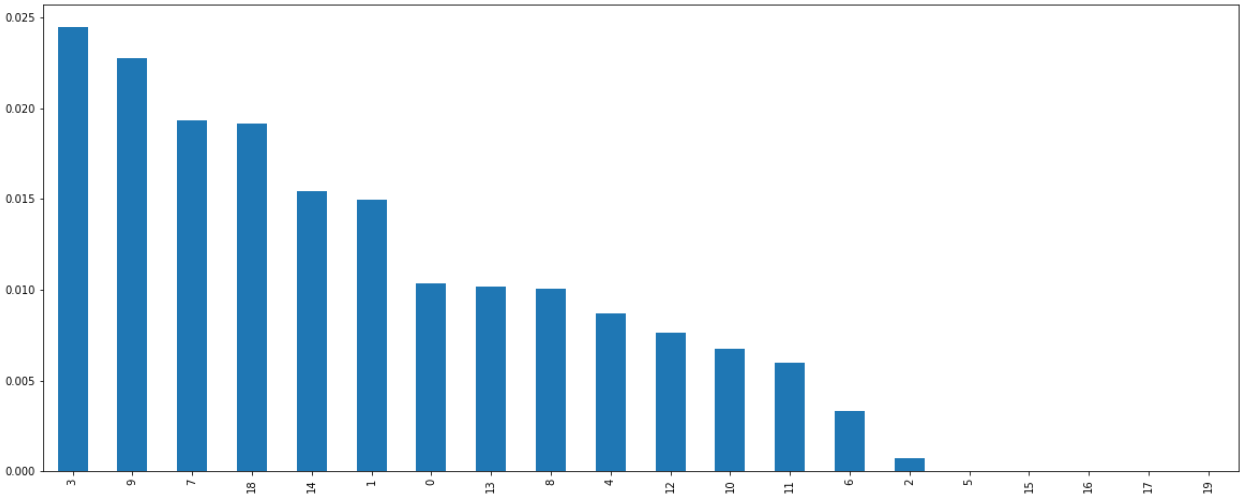


From the above chart we can infer that the features (except for the one hot encodings) are not much significantly correlated.

2. Using Mutual Information

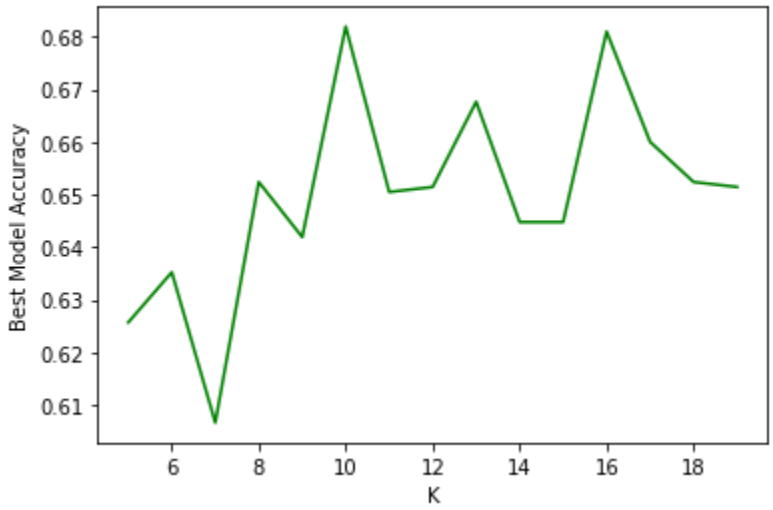
Mutual information is statistical method used to understand how a variable helps the model understand about the target variable.

It is different from Correlations!



3. Using KBEST

The kbest method picks k best features in the dataset. Following is the chart when I put k in range 5 to 20 and run. Surprisingly the model gives a different curve each time and hence I decided to go with all possible features.



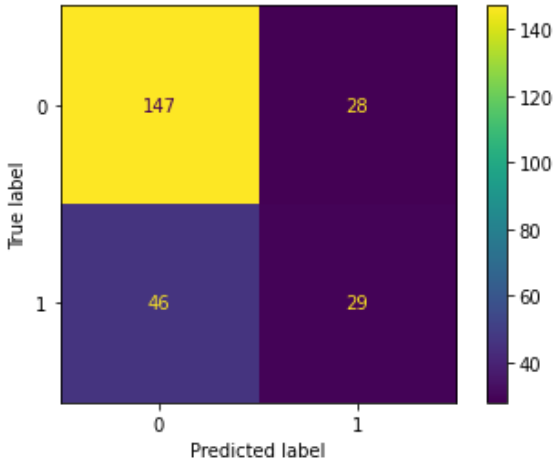
Model Building

I would like to mention few points before starting building a model

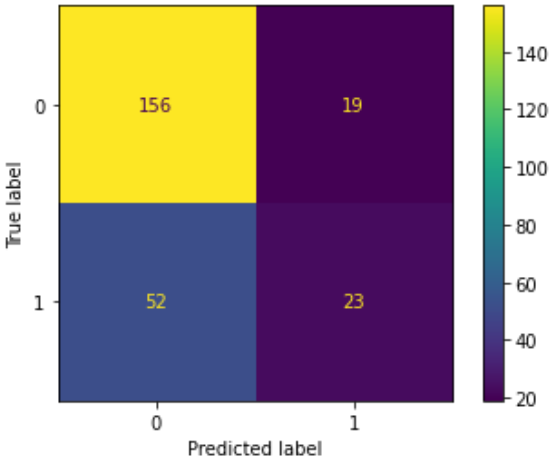
Business Constraint: Note that it is worse to state an applicant as a low credit risk when they are actually a high risk, than it is to state an applicant to be a high credit risk when they aren't.

- The above problem states that we should aim to minimize the False Positives or the Type 1 error should be Minimized!
- To minimize the Type, one errors or False Positives we must focus on models having more **Precision**
- Since the dataset is nearing an imbalance, it makes it necessary to stratify our splits. Hence, we will be performing Stratified Train, Test Splits while building the model
- Since, the dataset is nearing an imbalance and we also, do not have the sufficient data to under sample it. we will be **Oversampling** the data and train the model accordingly
- We will be building 2 models with imblearn library which can deal with imbalanced dataset.

Results of the 2 Models trained



Random Forest classifier



Logistic Regression Classifier

Hyper-Parameter Tuning on Random Forest Classifier

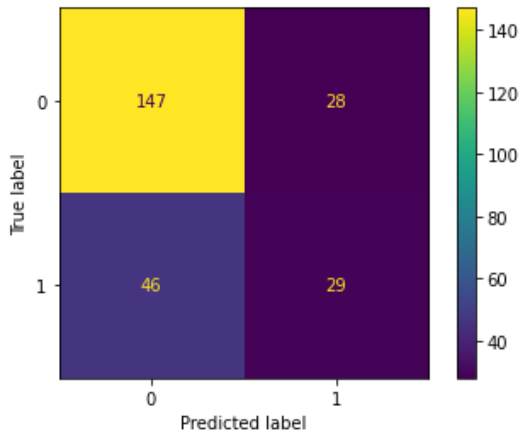
I have used randomSearchCV method for Hyperparameter tuning and found that the best Precision is at the following parameters

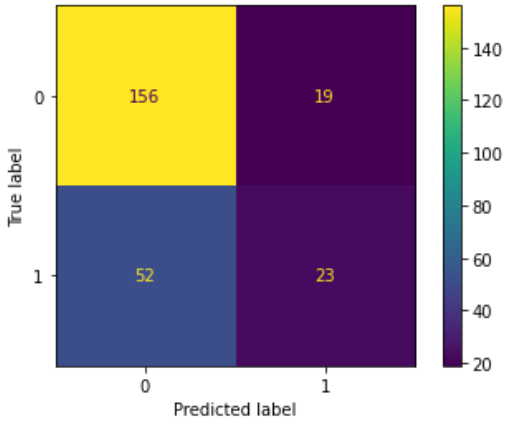
```
rf_randomcv.best_params_
```

```
{'n_estimators': 600,
 'min_samples_split': 5,
 'min_samples_leaf': 1,
 'max_features': 'sqrt',
 'max_depth': 780,
 'criterion': 'gini'}
```

Comparison of tuned and not tuned model

1. Confusion Matrix

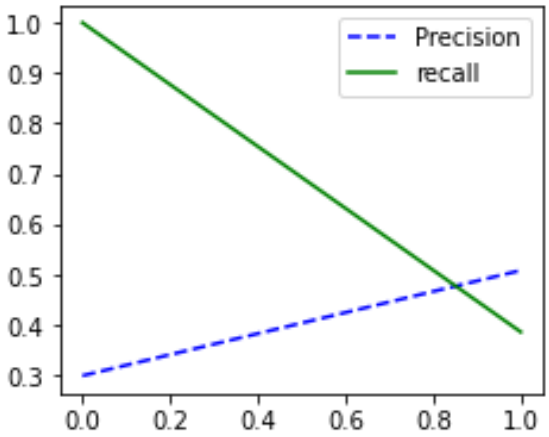


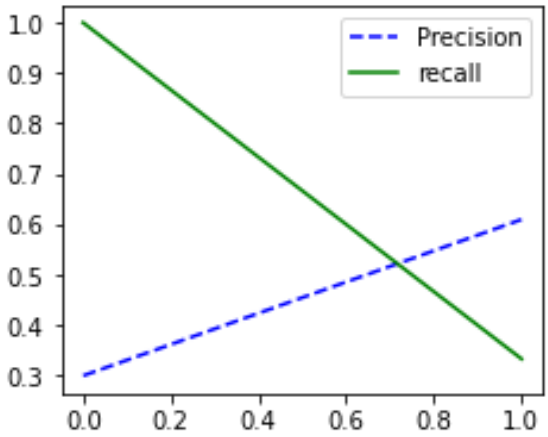


Not Tuned Tuned

You can see the type 1 error(upper right corner reduced from 28 to 19)

2. Precision Recall Curve





Not Tuned model Tuned model

Hence, we can observe the slope of precision increasing.

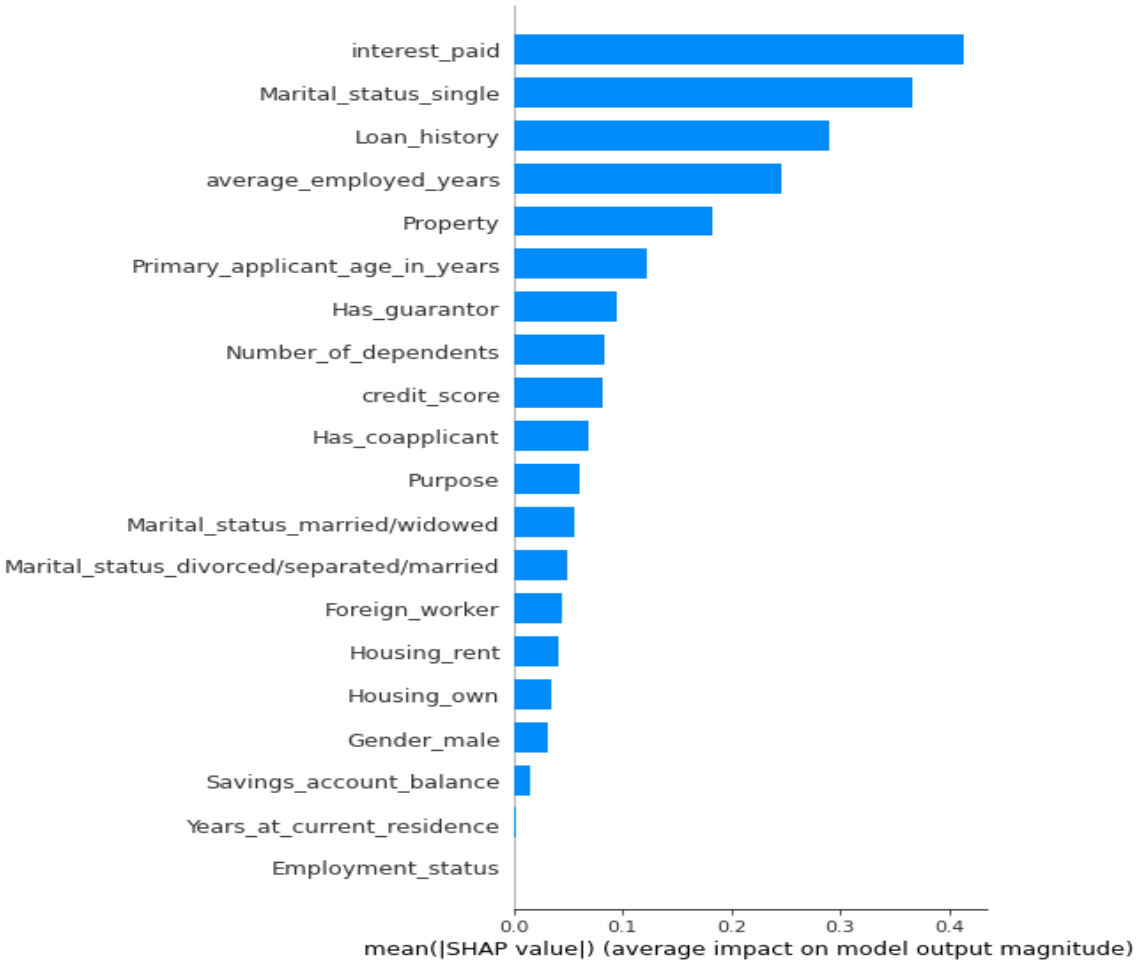
Explainable AI

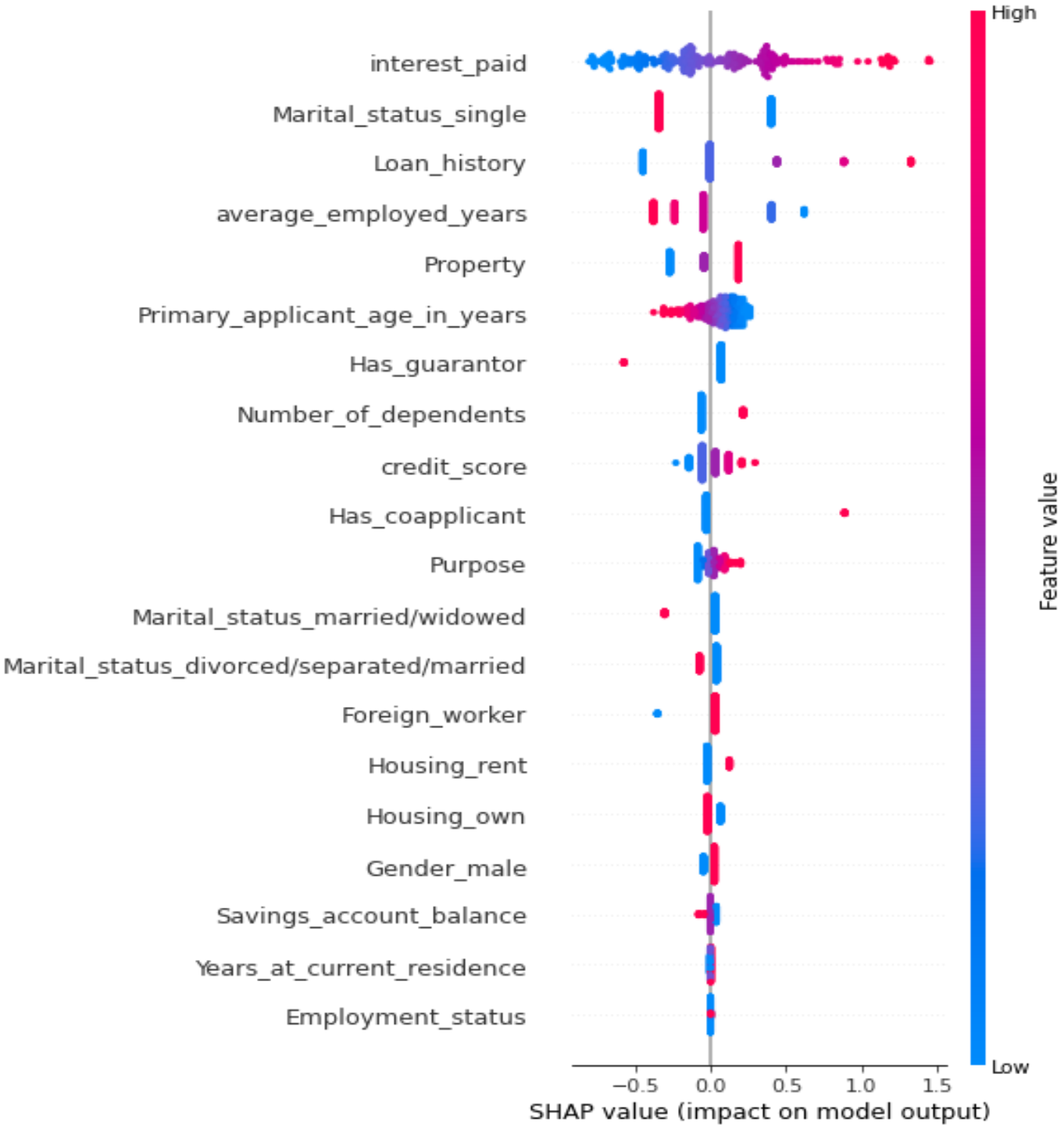
I have used the shap library to explain the models and their understandings.

The shap library has an inbuilt class called explainer, this explainer takes in the features and calculates an understanding value of the variable.

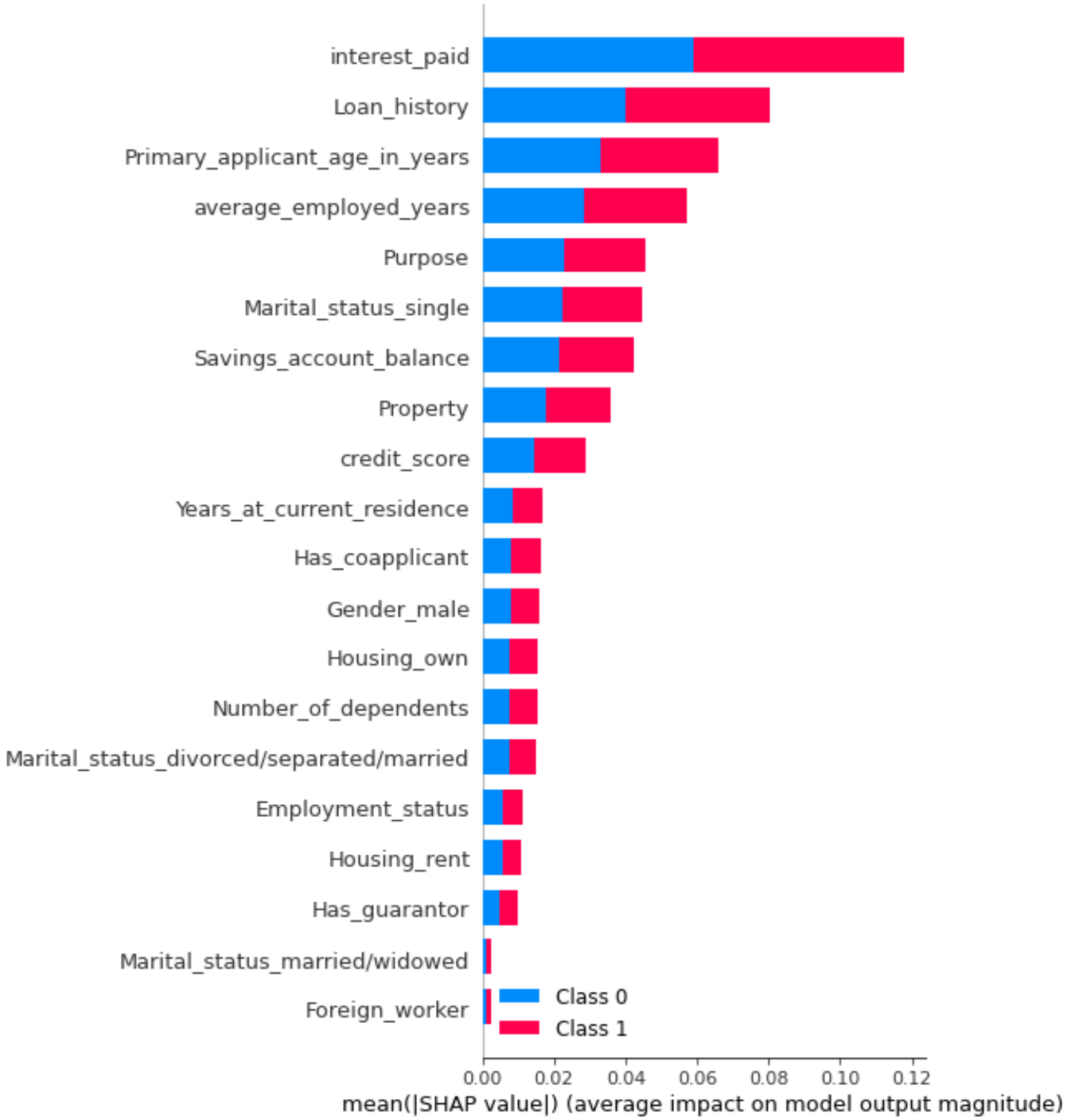
The understanding value is a measure of how the parameter helps the model to understand the problem statement and make better decision.

1. For logistic Regression model.




interest_paid
Marital_status_single
Loan_history
average_employed_years
Property
Primary_applicant_age_in_years
Has_guarantor
Number_of_dependents
credit_score
Has_coapplicant
Purpose
Marital_status_married/widowed
Marital_status_divorced/separated/married
Foreign_worker
Housing_rent
Housing_own
Gender_male
Savings_account_balance
Years_at_current_residence
Employment_status
−0.5
0.0
0.5
1.0
1.5
SHAP value (impact on model output)
High
Low
Feature value

For HyperTuned Random Forest Classifier



Some features that I would also, have liked to see..

1. **Financial Literacy –** The financial Literacy of the applicant
2. **Income of the person – I**ncome separated as family income and personal income.
3. **Cibil score –** The cibil score of the person.

THANKYOU