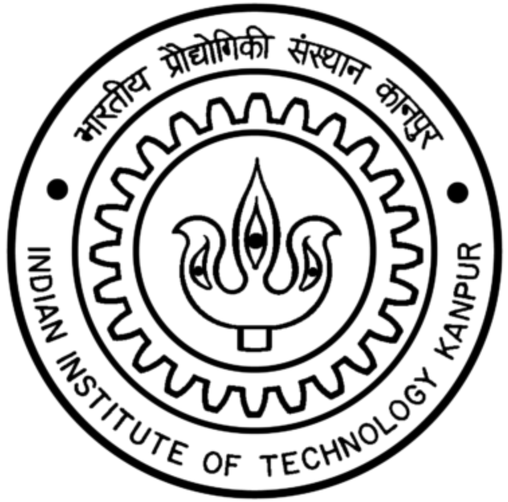


Under the Guidance of
Dr. Sharmishtha Mitra
Department Of Mathematics
And Statistics,
IIT Kanpur

• Loan Default Prediction by Logistic Regression

• *Submitted by* Souraj Mazumdar (211393), Soumya Paul (211391), Soumita Bandyopadhyay (211390) and Rahul Ghosh Dastidar (211353)

• *Date: 20.04.2022*

• Motivation

With the improvement of the banking sector in recent times and the increasing trend of loans, a large population asks for bank loans. But one of the major problem banking sectors face in this ever-changing economy is the increasing rate of loan defaults, and the banking authorities find it more difficult to properly assess loan requests and address the default risks of borrowers. The two most critical questions in the banking industry are (i) How risky is the borrower? and (ii) Given the borrower's risk, should we lend him/her?

• Objective of the Study

In this project, we wish to predict whether a customer is going to default his credit amount or not using Logistic Regression. The key steps which are involved in the process are as follows:

1. Dealing with the problem of Missing Values using suitable Data Imputation Techniques.
2. Building the model using Logistic Regression.
3. Checking the accuracy of the model.
4. Use the model for prediction purpose.

• Data Description

The Loan default dataset we have used in this study has been collected from Kaggle. The variables in the dataset are described as follows:

- **Id:** It refers to the Customer Id of the customer taking the loan.
- **Home Ownership:** Home ownership refers to the information on whether the home, in which the loan applicant is currently residing, is owned by him or it is rented or under mortgage .
- **Annual Income:**It refers to the total income of the customer during a financial year.
- **Years in current job:** It refers to the number of years the customer has been in his/her present job.
- **Tax Liens:** It refers to the no. of times a customer was penalized for failure of tax payment.
- **Number of Open Accounts:** The number of accounts (open) for a particular customer.

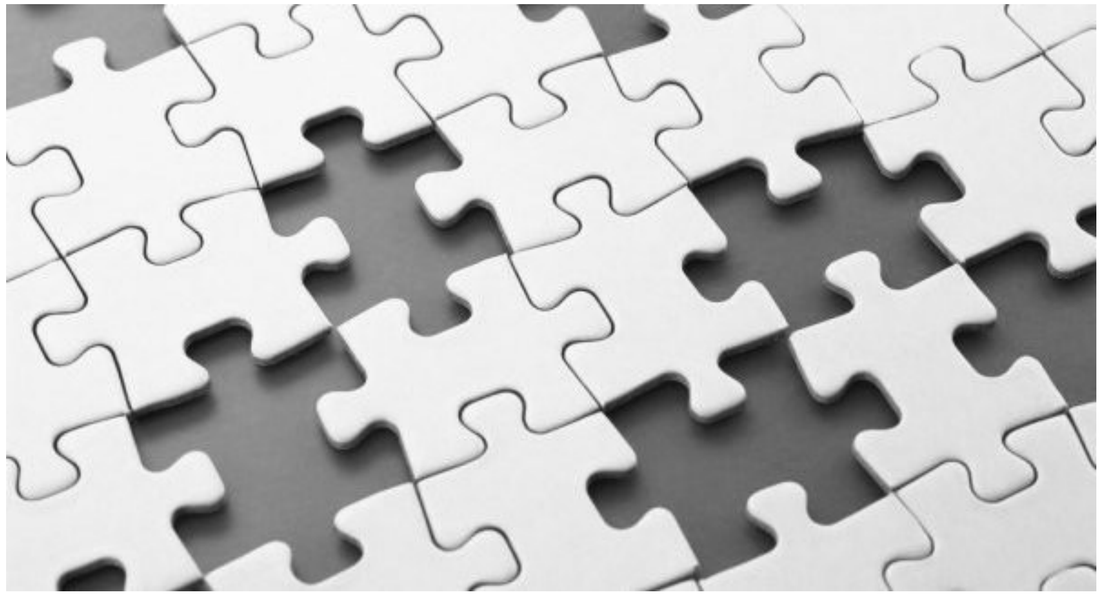
- **Years of Credit History:** The time span covering the issue of the first loan to the closure of the last loan of a customer.
- **Maximum Open Credit:** It is the maximum amount of credit available to the customer. The limit is revisable, and the borrower can request an increase in the maximum credit limit if the limit is not enough for their needs.
- **Number of Credit Problems:** The number of times a customer has experienced breakdown of his/her financial system caused by a sudden and severe disruption of the normal process of cash movement.
- **Months since last delinquent:** Months since last payment failure/delay.
- **Bankruptcies:** The number of times a customer has faced the issue of insolvency.
- **Purpose:** The purpose for taking the loan as stated by the customer.

- **Term:** The category of time period for which the loan has been taken - short term or long term.
- **Current Loan Amount:** The total amount of the ongoing loan the customer has taken.
- **Current Credit Balance:** The amount of credit which the customer is yet to pay back. It is basically Total Loan amount - Total amount that has been repayed.
- **Monthly Debt:** It refers to the equated monthly instalment (EMI), which are payments made regularly to repay an outstanding loan within a certain time frame.
- **Credit Score:** It is an indicator of a person's creditworthiness, or their ability to repay debt.
- **Credit Default:** A credit default occurs when a borrower is unable to make timely payments, misses payments, or avoids or stops making payments on interest or principal owed.

• Datasets in our study:

- Test dataset
- Train dataset

• Dealing with missing data



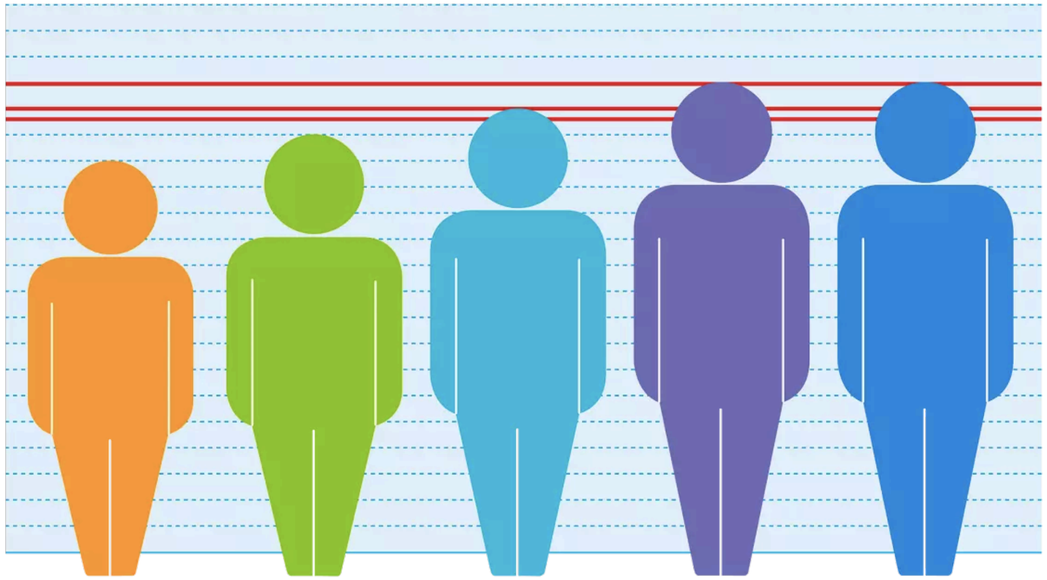
We use different data imputation techniques to deal with the problem of Missing Data -

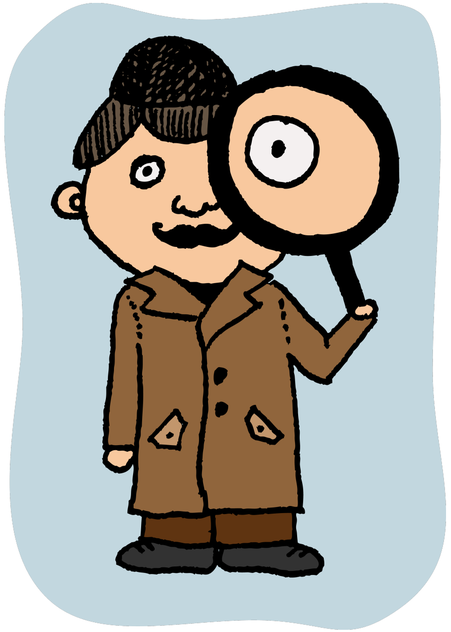
1. Mean Imputation

2. Median Imputation

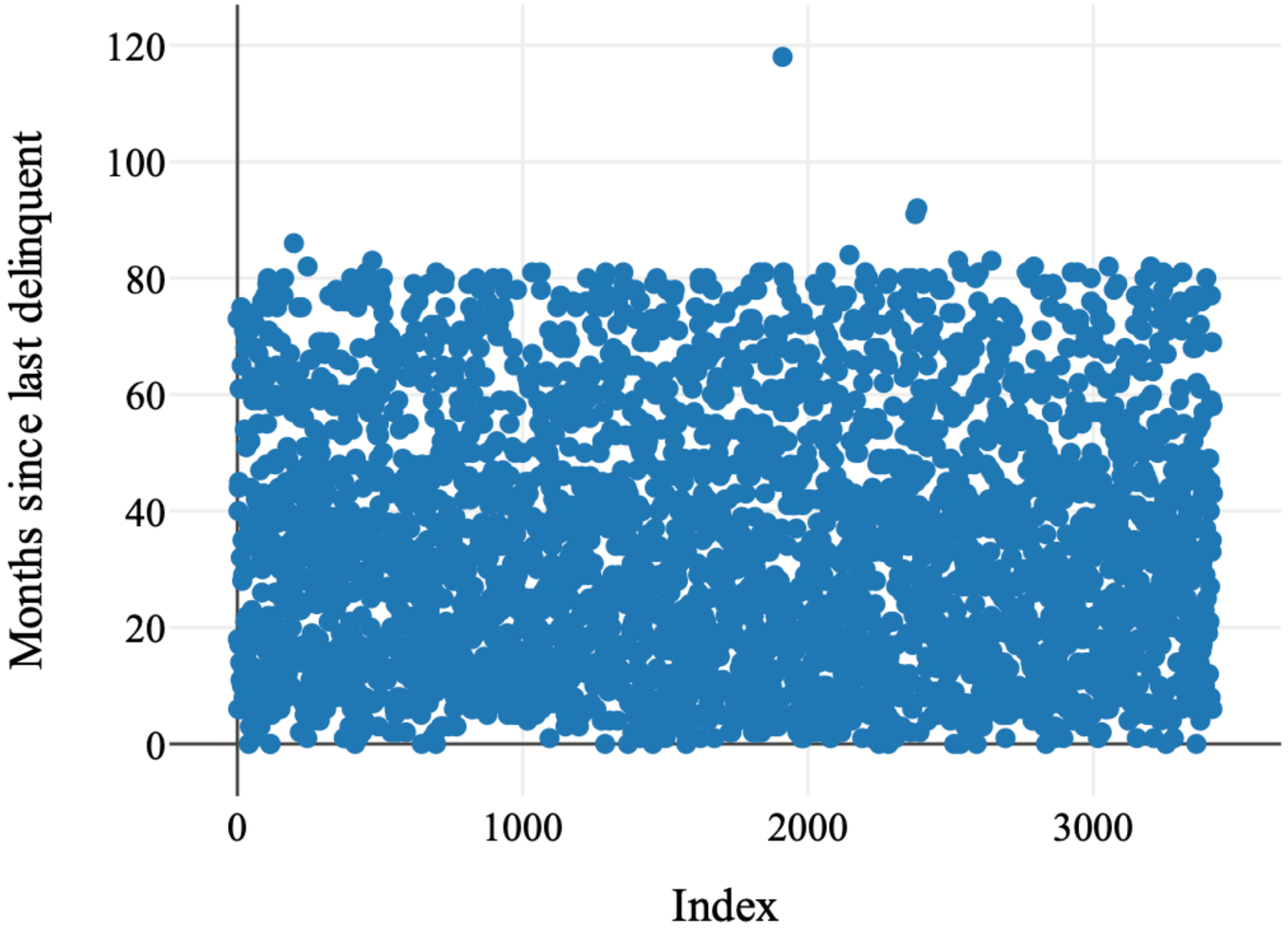
• Mean Imputation

- Mean imputation technique is the process of replacing any missing value in the data with the mean of that variable in context. In our dataset, we replaced missing values of a variable with the mean of the other non-missing values of that feature.



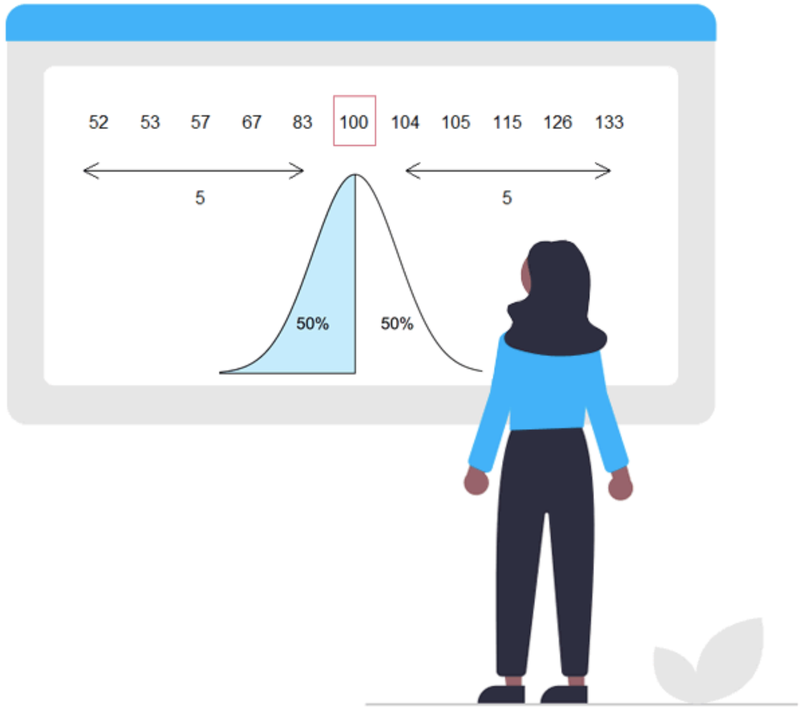


- From the plot of the variable Months since last delinquent, it is evident that mean imputation will be appropriate for this variable.

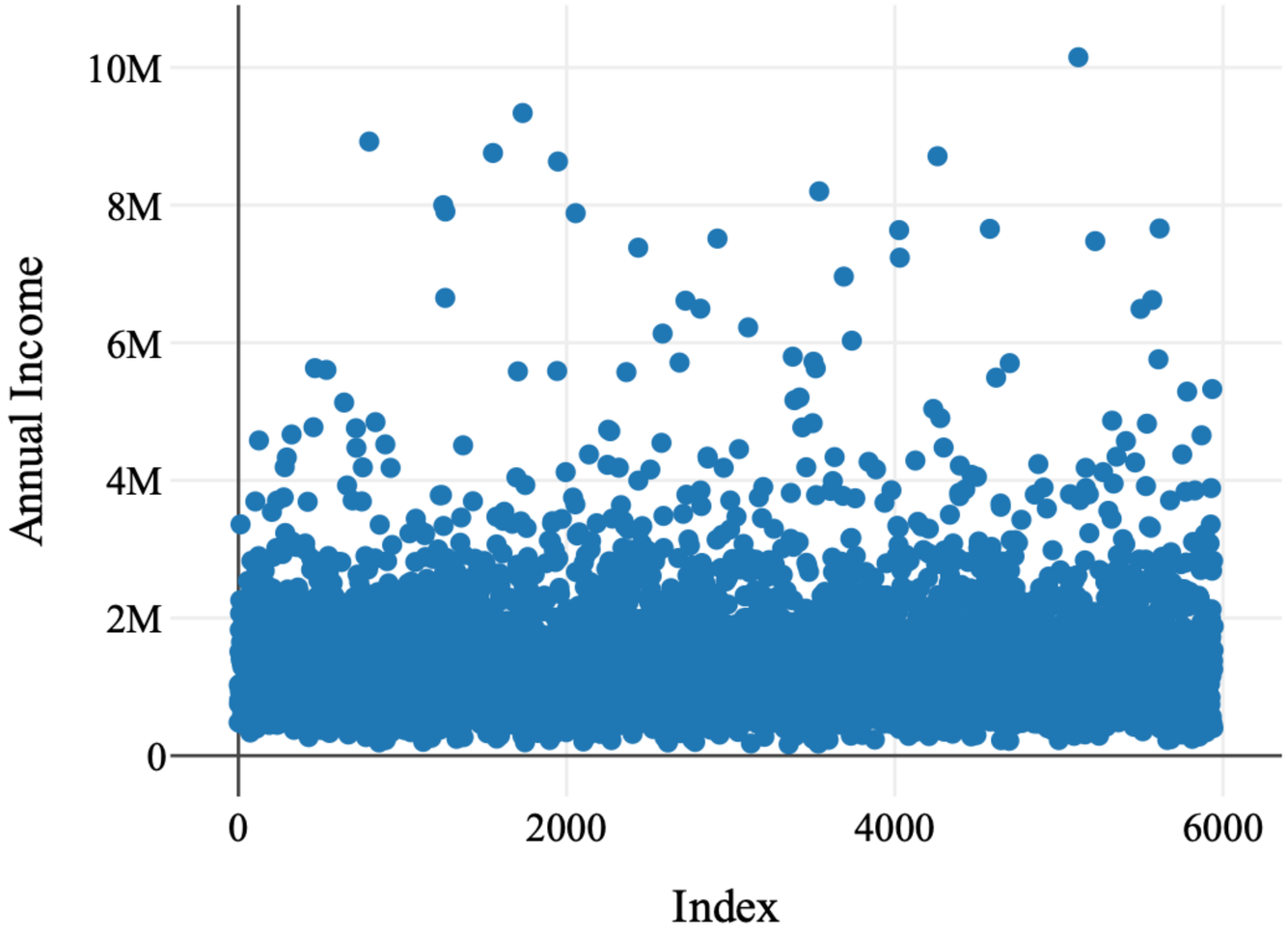


Median Imputation

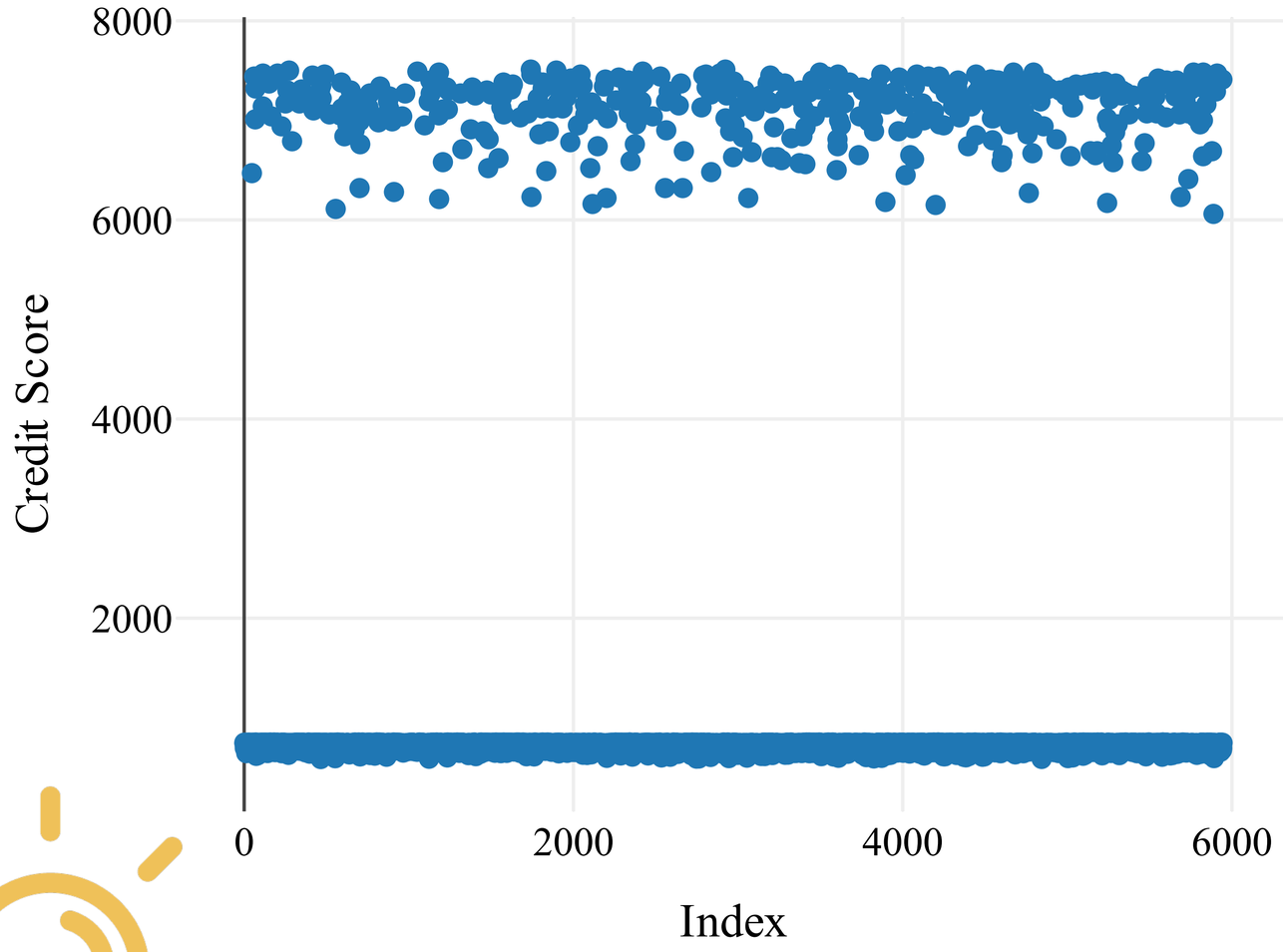
- In Median Imputation technique, one replaces the missing values with the median of the available values of the same variable. It is used mainly when the data is skewed. Also, median imputation technique is used when there is outliers in the data.

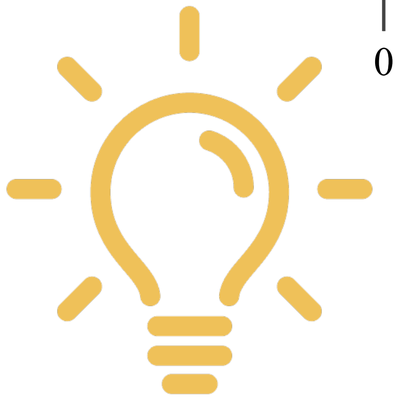


From the plot of the Annual Income, we can see that the observations are very dense for the lower income group whereas there are few observations for the higher income group. So, here we apply the median imputation.



- As we can see missing values of Credit score data are not very easy to impute
- Also there are 14 missing values for Bankruptcy variable



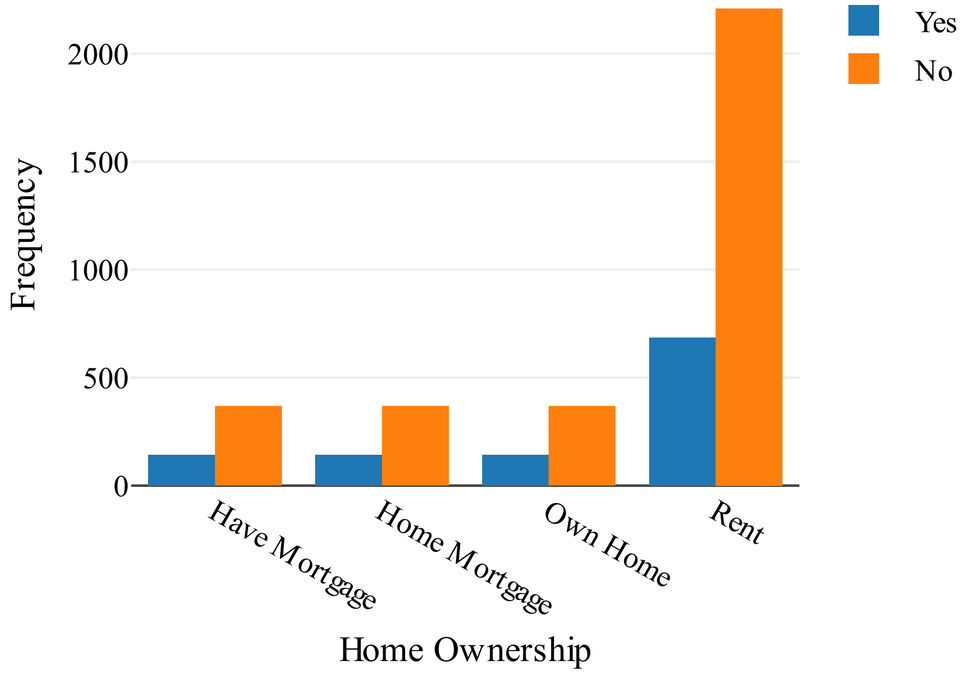


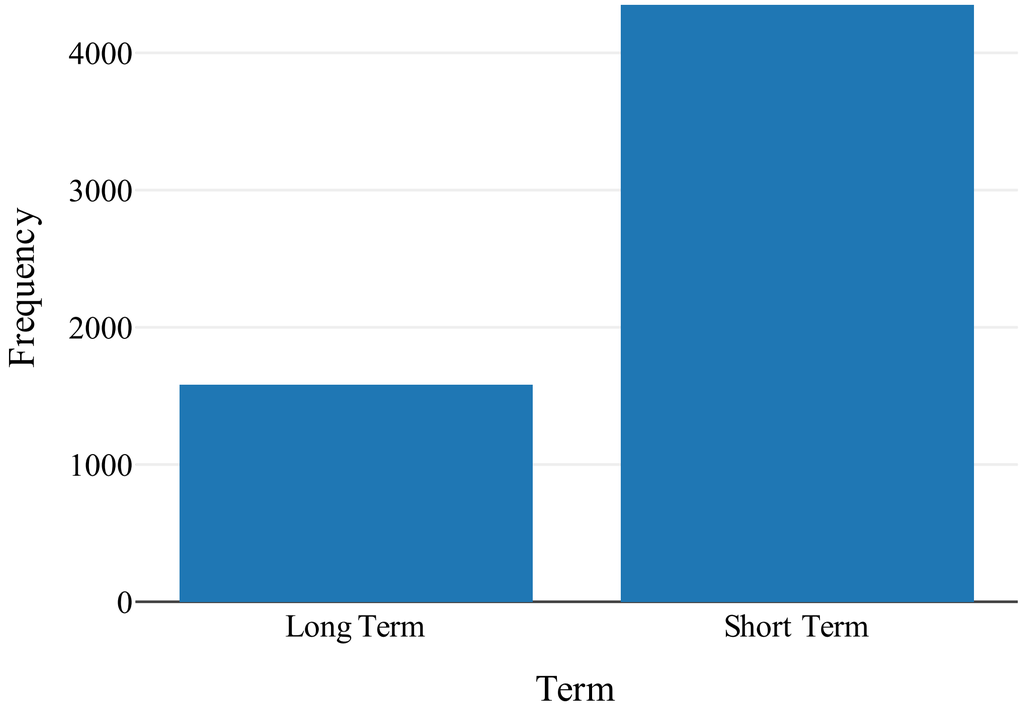
• Finally got data for further analysis!!

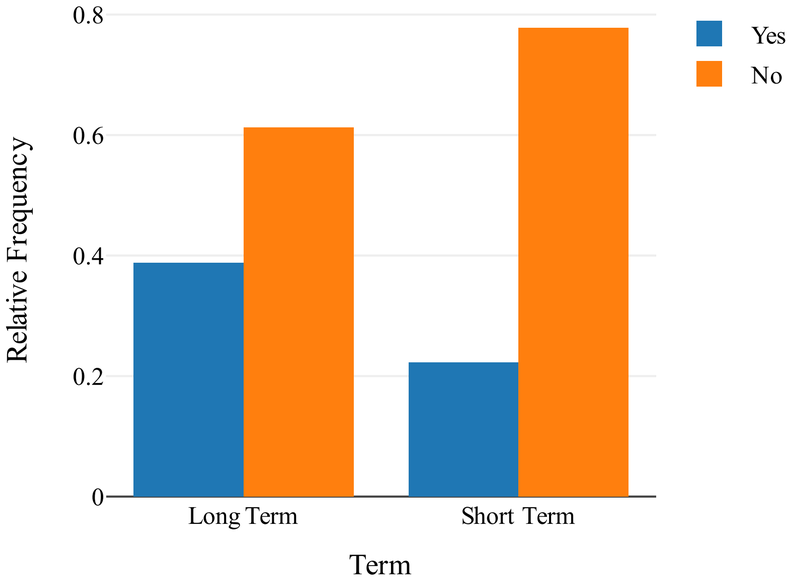
- After performing the mean and median imputation, we discard the missing values of the variables Bankruptcies and Credit Score
- So, now our data has been cleaned and we can proceed with our further analysis of this dataset.

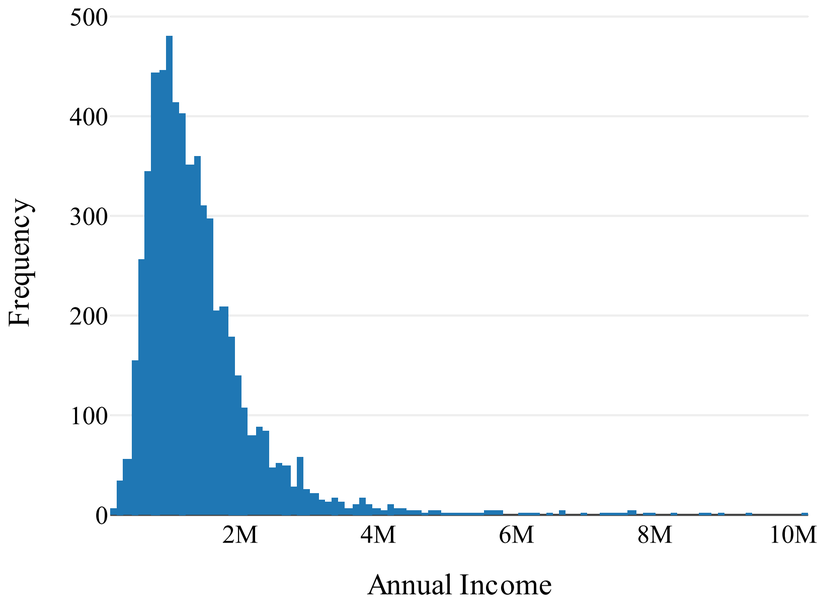
Exploratory data analysis

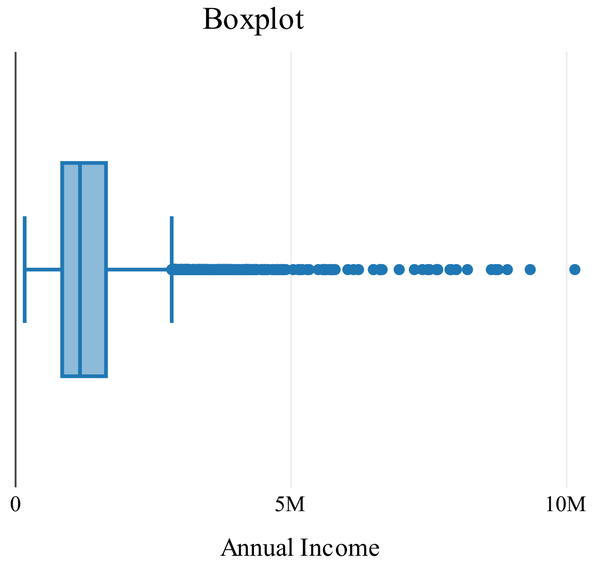
An exploratory data analysis is always helpful to get an insight about the data. So here we will try to visualise different variables by various plots and diagrams and try to analyse them.


Yes
No
2000
1500
1000
500
0
Frequency
Have Mortgage
Home Mortgage
Own Home
Rent
Home Ownership


4000
3000
2000
1000
0
Frequency
Long Term
Short Term
Term

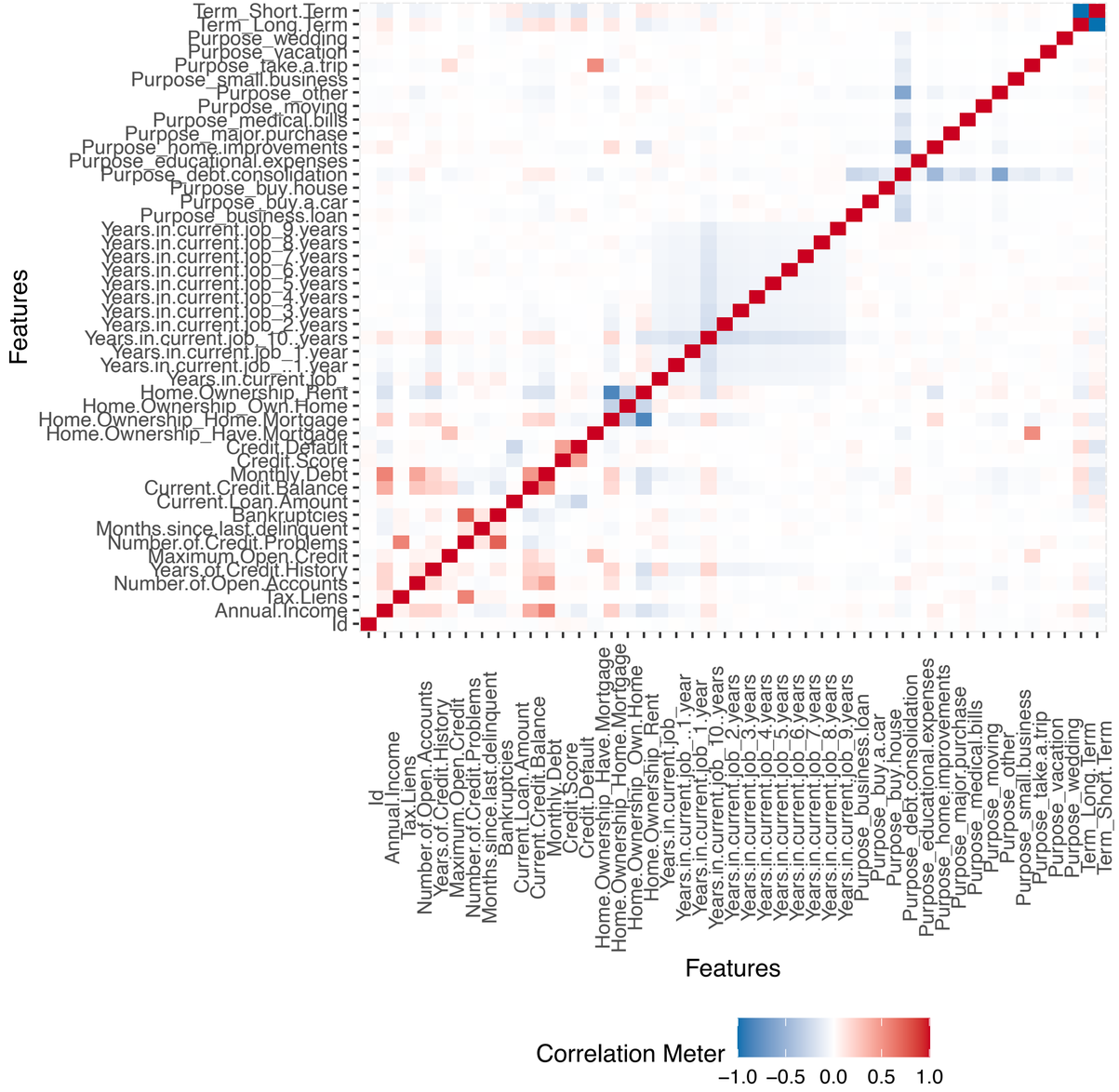

Yes
No
0.8
0.6
0.4
0.2
0
Relative Frequency
Long Term
Short Term
Term


500
400
300
200
100
0
Frequency
2M
4M
6M
8M
10M
Annual Income


Boxplot
0
5M
10M
Annual Income

• Correlation Heat map

- The following correlation heatmap is showing the association among the different variables of our train dataset.
- Formalising this mathematically, the definition of correlation usually used is Pearson's R for for continuous variables
- For discrete features, the function first dummifies all categories, then calculates the correlation matrix



Multicollinearity

- Several multicollinearity diagnostic measures are available. Here we have used "Variance Inflation Factor" to detect multicollinearity among the continuous variables of our dataset. The variance inflation factor for the jth explanatory variable (when all the regressors are scaled to unit norm) is defined as:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j denotes the coefficient of determination obtained when X_j is regressed on the remaining regressor variables.

- In practice, usually a $VIF > 5$ indicates that the corresponding explanatory variable is involved in multicollinearity.

So, we can see that our continuous regressors are free from multicollinearity


```
> VIF(model.cont)
                scale(Annual.Income)
                            1.836466
     scale(Years.of.Credit.History)
                            1.086024
         scale(Maximum.Open.Credit)
                            2.921941
scale(Months.since.last.delinquent)
                            1.007035
         scale(Current.Loan.Amount)
                            1.002746
      scale(Current.Credit.Balance)
                            3.284842
                scale(Monthly.Debt)
                            2.050633
                scale(Credit.Score)
                            1.012040
```

Variable Selection

- Variable selection is the method for selecting a subset of 'best' regressors from a pool of potential regressors.
- After checking multicollinearity, we are left with all the continuous regressors of our dataset. We also have other categorical regressors in our hand. Now, to follow the principle of parsimony, i.e. include as few as regressors as possible to explain the response variability in efficient manner, we go for variable selection.
- We use Forward Selection, Backward Selection and Stepwise Selection method for the aforesaid purpose. We compare the different AIC values of the different models obtained from the different variable selection methods and continue with that set of regressors which yield the minimum AIC value.

For full model

```
> log.reg$aic
[1] 5046.196
```

By forward selection

```
> step.model_for$aic
[1] 5046.196
```

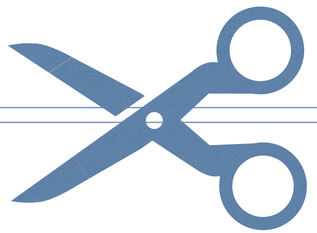
By backward selection

```
> step.model_back$aic
[1] 5039.023
```

By stepwise selection

```
> step.model_both$aic
[1] 5039.023
```

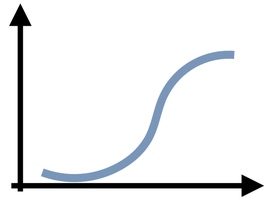
• Reduced Model



```
> step.model_both$call
glm(formula = Credit.Default ~ Home.Ownership + Annual.Income +
    Number.of.Open.Accounts + Years.of.Credit.History + Maximum.Open.Credit +
    Months.since.last.delinquent + Purpose + Term + Current.Loan.Amount +
    Current.Credit.Balance + Monthly.Debt + Credit.Score, family = binomial,
    data = train_data)
```

From the output we can see Home.Ownership ,Annual.Income, Number.of.Open.Accounts , Years.of.Credit.History , Maximum.Open.Credit ,Months.since.last.delinquent , Purpose , Term , Current.Loan.Amount ,Current.Credit.Balance , Monthly.Debt and Credit.Score can be used as regressors

Model : Logistic Regression



Now we are about to fit our Logistic Regression model. Our response variable Y is a categorical variable with only 2 categories, i.e. 0 (which indicates that the loan will not default) and 1(which indicates that the loan will default).

The logistic regression model is given by -

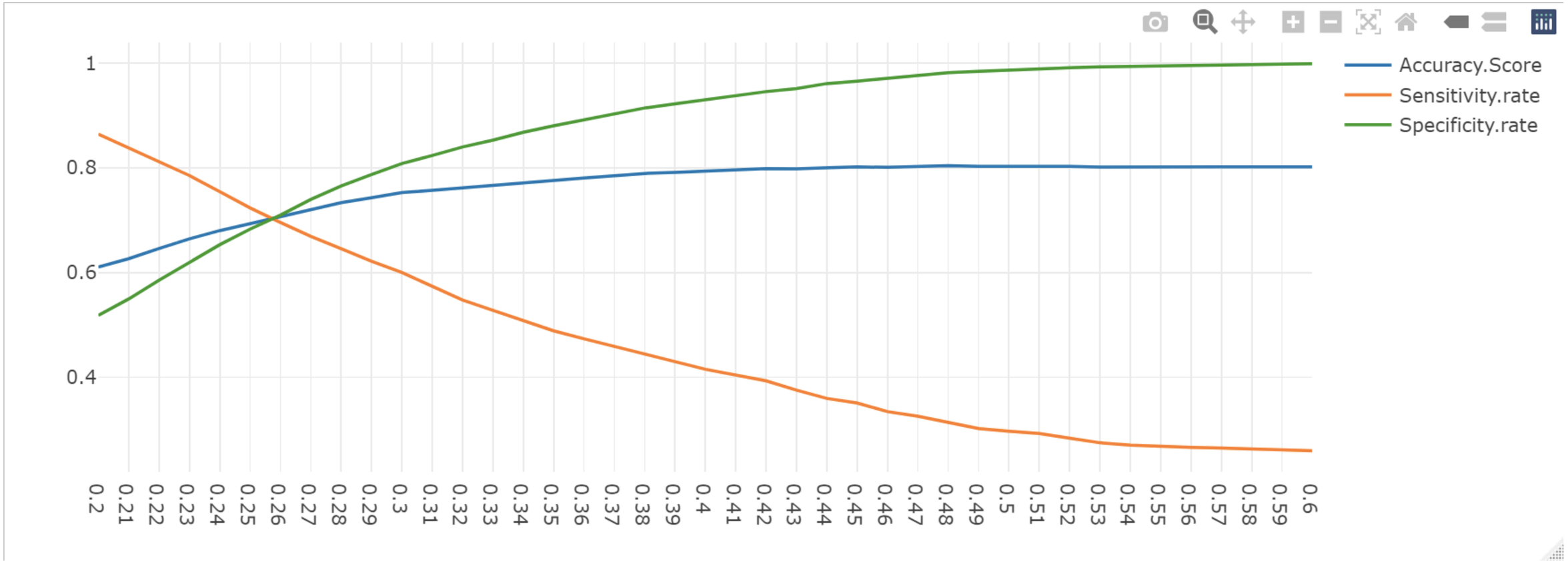
$$Y_i \sim Binomial(n_i, \pi_i)$$

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = \boldsymbol{X\beta} \quad \text{for } i = 1,2,.....,k$$

where x_{ij} is the element in the i^{th} row and $(j+1)^{th}$ column of the design matrix X. Here the regressors are qualitative and/or quantitative. Here the unknown probabilities $\pi_i's$ are estimated using Maximum Likelihood Method. Then we classify all $\widehat{Y_i} = 1, \quad if \quad \widehat{\pi_i} > c \quad and \quad 0 \quad if \quad \widehat{\pi_i} \leq c$ where c is a constant. In our model we will choose the value of c which will yield moderately high values of each of the measures - accuracy score, sensitivity rate and specificity rate (the rates are defined later)

Finding the Threshold Value c

The following graph shows the values of accuracy rate, specificity rate and sensitivity rate for different values of c. From the above plot we choose the value of c to be 0.3, which gives moderately high values of each of the measures - accuracy score,



• Interpretation of the fitted model

In general, β_j (estimated coefficient of the jth regressor) is the change in log-odds of $Y = 1$ for a unit change in x_j . In other words β_j is the change in odds ratio of $Y = 1$ for a unit change in x_j . Also the estimates which have less p-value, are more significant, and the estimates having larger p-value are less significant. As for example -

- 1 unit change in Annual Income will increase the odds of loan default by $\exp(-6.852 \times 10^{-7}) = 0.99$, keeping the other regressors fixed, and its p-value indicates that it is significant in determining the loan default.
- A Short Term loan is $exp(-8.970 \times 10^{-01}) = 0.4077912$times more likely to get default than a Long Term loan, keeping the other regressors fixed, and its p-value also indicates that it is significant in determining the loan default.

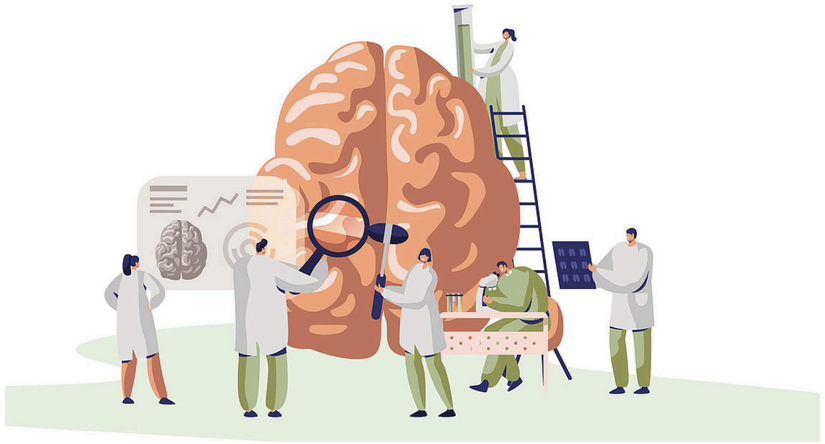
• Confusion Matrix

We got the following confusion matrix:

Predicted Value \ Observed Value	0	1
0	3551	670
1	800	910

• Some Diagnostics from the confusion matrix

- The higher the accuracy rate, the better is the model. For our model, the accuracy score is 75.21%
- The higher the sensitivity rate, the better is the model. For our model, the sensitivity rate is 57.59%
- The higher the specificity rate, the better is the model. For our model, the specificity rate is 81.61%



• Some more diagnostics

- Deviance test:

  For our model, the deviance statistic is less than the critical value of the chi-square distribution with corresponding degrees of freedom. Hence, we conclude that the fitted model is close to the saturated model at 5% level of significance.

- Pearson's Chi-square test:

  The observed value of the chi-square statistic is greater than critical value of the chi-square distribution with corresponding degrees of freedom. From here we conclude that Y and $\hat{Y}$ are associated.

• Some more diagnostics

- Phi- coefficient:

  The higher value of ϕ indicates stronger association between Y and $\hat{Y}$. Our observed value 0.383 signifies a moderate association.

- Contingency Coefficient:

  The higher value of P indicates stronger association between Y and $\hat{Y}$. Our observed value 0.3570 signifies a moderate association.

Cleaning of test data

- We have used the same methodology that we applied for the cleaning of train dataset as per requirement, on the test dataset.
- We then proceeded with the prediction for test data.

• Prediction of Credit Default for test data

- Here we applied the same model to predict whether the customers present in the test dataset will default his/her loan or not.
- Using R, we generate the required Credit default values on the test data.