Project: Predicting Loan Defaulters

By: Vaibhav Bajaj

Background:

Financial institutions incur significant losses due to the default of vehicle loans. This has led to the tightening up of vehicle loan underwriting and increased the vehicle loan rejection rates. The need for a better credit risk scoring model among these institutions also gets created. This warrants a study to estimate the determinants of vehicle loan default.

Importing, Understanding, and Inspecting Data :

i. **Perform preliminary data inspection and report the findings as the structure of the data, missing values, duplicates, etc.**

Solution:

We can use the summary function and get the details for all the variables in the data frame created. This will help us understand the data by knowing if it has any missing values or outliers etc.

Code:

```
summary(Loan)
```

Outcome with Screenshot of the console:

I. UniqueId is a unique identifier for customers and thus it will not affect the loan defaulting in any way. Thus, it can be dropped for model creation.

II. disbursed_amount and asset_cost might have an outlier present as can be seen by comparing the median and mean of each case separately. Thus there might be a need for outlier removal there.

III. Branch_id, supplier_id, manufacturer_id, current_pincode_id, state_Id, Employee_code_id are all numeric variables. Maybe we should convert them to factors to make use in the model as we cannot ignore them as there might be a case of fraudulence that is being conducted in a specific branch or manufacturer or there an area might be notoriously having a large number of defaults, thus maybe the bank would want to avoid giving loans there. (A little far-fetched idea, but it seemed logical to keep here.)

Note : Not converting them to factors due to memory limits of R not able to create a logistic regression model with so many factors. Although they can be converted if any other algorithm is used while model creation.

IV. Employment.type is a character vector. It should also be converted to factor. There were many empty cells under this field in the data excel sheet. When this Excel sheet was imported to an R environment, these cells were marked as N/As.

There are two approaches here:

1) We can convert these into a third employment type and store accordingly.
2) We can omit all those observations where employment.type is N/A.

Note: During model creation which is discussed in this document, the 2nd approach is followed.

V. Date.of.Birth and DisbursalDate can be used to calculate the age of the customer at the time of disbursal of loan amount. Thus, these two variables can be replaced by a single variable as Age which can be numeric or can convert to factor depending on the model employed.

VI. MobileNo_Avl_Flag has the minimum and maximum values as 1. This means that it is not affecting the loan_default in any way.

VII. Adhar_flag, PAN_flag, VoterId_flag, Driving_flag, Passport_flag are all numeric vectors. We can convert them to factors.

Note: No factor conversion done here too for the logistic regression model due to the R memory usage limit.

VIII. PERFORM_CNS.SCORE has a huge difference in the values of mean and median. Outliers might be present there and thus, outlier removal might be required here. But there might be a another way where we might not even require this variable as explained in the next point.

IX. PERFORM_CNS.SCORE.DESCRIPTION is the category assigned to PERFORM_CNS.SCORE. Thus we can just convert this field into factor and remove the variable PERFORM_CNS.SCORE from the model entirely as the corelation between these two variables would be very high. That way we would not have to work with outlier removal as well.

Note: This approach is not followed during the creation of final model discussed in this document.

X. PRI.NO.OF.ACCOUNTS might require some outlier removal as the values of mean and median are different. Although since the values are so close the need might not be there.

Note: This step of outlier removal was not done as scaling was performed and it took care of most of the outliers in the data related to this variable.

XI. Same for PRI.ACTIVE.ACCTS, PRI.OVERDUE.ACCTS.

XII. PRI.CURRENT.BALANCE is total Principal outstanding amount of the active loans at the time of disbursement. The minimum value for this should be 0. Thus, will need to change all the values that are below 0 to 0.

XIII. PRI.SANCTIONED.AMOUNT, PRI.DISBURSED.AMOUNT has a huge difference in the values of mean and median. Thus, presence of outlier is possible. Outlier removal might be required here.

XIV. SEC.CURRENT.BALANCE may have an outlier as the differenece is huge in median and mean. Also, the minimum value of this variable can be 0. Thus all the values that are below 0 needs to be converted to 0.

XV. SEC.SANCTIONED.AMOUNT, SEC.DISBURSED.AMOUNT, PRIMARY.INSTAL.AMT, SEC.INSTAL.AMT may require outlier removals as well as seen by the difference in mean and median.

XVI. NEW.ACCTS.IN.LAST.SIX.MONTHS might have an outlier but the chances are very slim. Might try just to be on the safe side.

XVII. AVERAGE.ACCT.AGE, CREDIT.HISTORY.LENGTH are character vectors. They need to be converted to numeric vectors as they give the duration of an average account held by the customer and also the duration of credit history given by the customer.

XVIII. NO.OF_INQUIRIES may or may not be an important variable at all as it is just the number of inquiries made by a customer for a loan. Will skip this in the initial model but include it in the next models if the initial model is not as good as required.

XIX. Convert loan_default to factor values. This will help in easily knowing how many have defaulted and how many have not.

Note: Many variables are continuous numeric with greatly differing ranges. Thus they need to be scaled as well.

These are the following variables:

a) Age: of the customer at the time of disbursement of loan. This will be created before data modelling and will replace the Date of birth and Disbursal date.
b) If the PERFORM_CNS.SCORE variable is used instead of PERFORM_CNS.SCORE.DESCRIPTION variable, then it will need to be scaled as well.
c) PRI.NO.OF.ACCOUNTS, PRI.CURRENT.BALANCE and other variables related to the loans taken previously by the customer both as primary and secondary.

```
   UniqueID      disbursed_amount   asset_cost           ltv          branch_id    
 Min.   :417428   Min.   : 13320   Min.   :  37000   Min.   :10.03   Min.   :  1.00  
 1st Qu.:476786   1st Qu.: 47145   1st Qu.:  65717   1st Qu.:68.88   1st Qu.: 14.00  
 Median :535979   Median : 53803   Median :  70946   Median :76.80   Median : 61.00  
 Mean   :535918   Mean   : 54357   Mean   :  75865   Mean   :74.75   Mean   : 72.94  
 3rd Qu.:595040   3rd Qu.: 60413   3rd Qu.:  79202   3rd Qu.:83.67   3rd Qu.:130.00  
 Max.   :671084   Max.   :990572   Max.   :1628992   Max.   :95.00   Max.   :261.00  
  supplier_id    manufacturer_id  Current_pincode_ID Date.of.Birth                  
 Min.   :10524   Min.   : 45.00   Min.   :   1       Min.   :1949-09-15 00:00:00  
 1st Qu.:16535   1st Qu.: 48.00   1st Qu.:1511       1st Qu.:1977-05-04 00:00:00  
 Median :20333   Median : 86.00   Median :2970       Median :1986-01-01 00:00:00  
 Mean   :19639   Mean   : 69.03   Mean   :3397       Mean   :1984-04-04 04:32:39  
 3rd Qu.:23000   3rd Qu.: 86.00   3rd Qu.:5677       3rd Qu.:1992-05-19 00:00:00  
 Max.   :24803   Max.   :156.00   Max.   :7345       Max.   :2000-10-20 00:00:00  
 Employment.Type     DisbursalDate                    State_ID      Employee_code_ID
 Length:233154      Min.   :2018-08-01 00:00:00   Min.   : 1.000   Min.   :   1    
 Class :character   1st Qu.:2018-08-30 00:00:00   1st Qu.: 4.000   1st Qu.: 713    
 Mode  :character   Median :2018-09-25 00:00:00   Median : 6.000   Median :1451    
                    Mean   :2018-09-23 09:57:53   Mean   : 7.262   Mean   :1549    
                    3rd Qu.:2018-10-21 00:00:00   3rd Qu.:10.000   3rd Qu.:2362    
                    Max.   :2018-10-31 00:00:00   Max.   :22.000   Max.   :3795    
 MobileNo_Avl_Flag  Aadhar_flag        PAN_flag         VoterID_flag   
 Min.   :1         Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
 1st Qu.:1         1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:0.0000  
 Median :1         Median :1.0000   Median :0.00000   Median :0.0000  
 Mean   :1         Mean   :0.8403   Mean   :0.07558   Mean   :0.1449  
 3rd Qu.:1         3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000  
 Max.   :1         Max.   :1.0000   Max.   :1.00000   Max.   :1.0000  
  Driving_flag      Passport_flag       PERFORM_CNS.SCORE PERFORM_CNS.SCORE.DESCRIPTION
 Min.   :0.00000   Min.   :0.000000   Min.   :  0.0     Length:233154                
 1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:  0.0     Class :character             
 Median :0.00000   Median :0.000000   Median :  0.0     Mode  :character             
 Mean   :0.02324   Mean   :0.002127   Mean   :289.5                                  
 3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:678.0                                  
 Max.   :1.00000   Max.   :1.000000   Max.   :890.0                                  
 PRI.NO.OF.ACCTS    PRI.ACTIVE.ACCTS PRI.OVERDUE.ACCTS PRI.CURRENT.BALANCE
 Min.   :  0.000   Min.   :  0.00   Min.   : 0.0000   Min.   :-6678296   
 1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.: 0.0000   1st Qu.:       0   
 Median :  0.000   Median :  0.00   Median : 0.0000   Median :       0   
 Mean   :  2.441   Mean   :  1.04   Mean   : 0.1565   Mean   :  165900   
 3rd Qu.:  3.000   3rd Qu.:  1.00   3rd Qu.: 0.0000   3rd Qu.:   35006   
 Max.   :453.000   Max.   :144.00   Max.   :25.0000   Max.   :96524920   
 PRI.SANCTIONED.AMOUNT PRI.DISBURSED.AMOUNT SEC.NO.OF.ACCTS     SEC.ACTIVE.ACCTS  
 Min.   :0.000e+00     Min.   :0.000e+00    Min.   : 0.00000   Min.   : 0.0000  
 1st Qu.:0.000e+00     1st Qu.:0.000e+00    1st Qu.: 0.00000   1st Qu.: 0.0000  
 Median :0.000e+00     Median :0.000e+00    Median : 0.00000   Median : 0.0000  
 Mean   :2.185e+05     Mean   :2.181e+05    Mean   : 0.05908   Mean   : 0.0277  
 3rd Qu.:6.250e+04     3rd Qu.:6.080e+04    3rd Qu.: 0.00000   3rd Qu.: 0.0000  
 Max.   :1.000e+09     Max.   :1.000e+09    Max.   :52.00000   Max.   :36.0000  
 SEC.OVERDUE.ACCTS   SEC.CURRENT.BALANCE SEC.SANCTIONED.AMOUNT SEC.DISBURSED.AMOUNT
 Min.   :0.000000   Min.   : -574647    Min.   :       0      Min.   :       0    
 1st Qu.:0.000000   1st Qu.:       0    1st Qu.:       0      1st Qu.:       0    
 Median :0.000000   Median :       0    Median :       0      Median :       0    
 Mean   :0.007244   Mean   :    5428    Mean   :    7296      Mean   :    7180    
 3rd Qu.:0.000000   3rd Qu.:       0    3rd Qu.:       0      3rd Qu.:       0    
 Max.   :8.000000   Max.   :36032852    Max.   :30000000      Max.   :30000000    
 PRIMARY.INSTAL.AMT SEC.INSTAL.AMT    NEW.ACCTS.IN.LAST.SIX.MONTHS
 Min.   :       0   Min.   :      0   Min.   : 0.0000             
 1st Qu.:       0   1st Qu.:      0   1st Qu.: 0.0000             
 Median :       0   Median :      0   Median : 0.0000             
 Mean   :   13105   Mean   :    323   Mean   : 0.3818             
 3rd Qu.:    1999   3rd Qu.:      0   3rd Qu.: 0.0000             
 Max.   :25642806   Max.   :4170901   Max.   :35.0000             
 DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS AVERAGE.ACCT.AGE   CREDIT.HISTORY.LENGTH
 Min.   : 0.00000                    Length:233154      Length:233154        
 1st Qu.: 0.00000                    Class :character   Class :character     
 Median : 0.00000                    Mode  :character   Mode  :character     
 Mean   : 0.09748                                                            
 3rd Qu.: 0.00000                                                            
 Max.   :20.00000                                                            
 NO.OF_INQUIRIES     loan_default   
 Min.   : 0.0000   Min.   :0.0000  
 1st Qu.: 0.0000   1st Qu.:0.0000  
 Median : 0.0000   Median :0.0000  
 Mean   : 0.2066   Mean   :0.2171  
 3rd Qu.: 0.0000   3rd Qu.:0.0000  
 Max.   :36.0000   Max.   :1.0000  
```

ii. **Variable names in the data may not be in accordance with the identifier naming in Python so, change the variable names accordingly**

Solution:

This project was created in R. Thus, no problem was faced regarding the data variable names as such. Thus skipping this step.

iii. **The presented data might also contain some missing values therefore, exploration will also lead to devising strategies to fill in the missing values while exploring the data**

Solution:

Missing values were found in the variable Employment.Type. Thus, there are 2 approaches that can be employed here leading to 2 different model creations:

A. Removing all the values that are missing.

Code:

Loan = na.omit(Loan)

B. Replacing the missing values with "Unknown".

Code:

Loan$Employment.Type = as.factor(Loan$Employment.Type)
Loan$Employment.Type=
ifelse(is.na(Loan$Employment.Type),"Unknown",Loan$Employment.Type)

Performing EDA and Modelling:

iv. **Provide the statistical description of the quantitative data variables**

Discussed above in point (i) in detail.

v. **Explain how is the target variable distributed overall**

Solution:

The target variable in this Project is **loan_default**. First we would need to convert this variable in factor as this is initially a numeric vector. This will make it easy for further modelling as well. Although the number of default loans can be calculated just by taking the sum of the variable as the values are in 0(not defaulted) and 1(defaulted). But since we have to convert this variable in factor nonetheless, we may as well do it in this step already.

Code:

```
#Not performing on orginal df to avoid loading it again and again.
Loan1 = Loan
#Saving the variable loan_default under Loan1 as factor.
Loan1$loan_default = as.factor(Loan1$loan_default)
#Gives the summary of the variable loan_default
summary(Loan1$loan_default)
```

Outcome:

```
> summary(Loan1$loan_default)
     0      1
182543  50611
>
```

As shown in the above screenshot, out of 233152 observations, 50611 are the ones in which the customer defaulted in repaying loan.

vi. **Study the distribution of the target variable across various categories like branch, city, state, branch, supplier, manufacturer, etc.**

Solution:

Created a Tableau Public Workbook for each variable. Each workbook includes the bar graph, a pareto giving %age of Loan Defaults against %age of each of those variables, and a cross tab.

a) BranchID vs Loan Defaults

https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/BranchvsLoanDefault/BranchvsLoanDefault

As seen in the image below, 20.73% branches are having 52.47% of defaults:

Branch vs Loan Default - CrossTab

Branch Id	% of Loan Defaults	%age of Branches in which default occurs	Loan Default
36	5.18%	1.22%	2,621
2	10.03%	2.44%	2,455
67	14.37%	3.66%	2,198
5	18.42%	4.88%	2,047
16	22.00%	6.10%	1,815
136	25.40%	7.32%	1,721
3	28.59%	8.54%	1,614
146	31.59%	9.76%	
34	34.40%	10.98%	
251	37.01%	12.20%	
18	39.32%	13.41%	1,169
74	41.63%	14.63%	1,168
10	43.88%	15.85%	1,142
147	46.11%	17.07%	1,129
120	48.32%	18.29%	1,118
61	50.52%	19.51%	1,111
65	52.47%	20.73%	989
11	54.33%	21.95%	942
19	56.18%	23.17%	933
48	58.02%	24.39%	931
138	59.81%	25.61%	905
20	61.51%	26.83%	865
1	63.20%	28.05%	853
79	64.68%	29.27%	751
103	66.12%	30.49%	727

Branch Id: 3
Loan Default: 1,614

b) Current Pincode Id vs Loan Defaults

https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/CurrentPincodeIdvsLoanDefaults/CurrentpincodevsLoanDefault-Pareto?publish=yes

As seen in the image below, 74.26% of loan defaults occur in 20.02% of Current Pincode Ids.

Curr..	% of Current pincode ID	% of Loan Default	Loan Default
1369	19.72%	73.89%	10.0
1523	19.74%	73.91%	10.0
1533	19.75%	73.93%	10.0
1545	19.77%	73.95%	10.0
1601	19.78%	73.97%	10.0
1608	19.80%	73.99%	10.0
1614	19.81%	74.01%	10.0
1637	19.83%	74.03%	10.0
1642	19.84%	74.05%	10.0
1643	19.86%	74.06%	10.0
1647	19.87%	74.08%	10.0
1648	19.89%	74.10%	10.0
1697	19.90%	74.12%	10.0
1705	19.92%	74.14%	10.0
1750	19.93%	74.16%	10.0
1759	19.95%	74.18%	10.0
1777	19.96%	74.20%	10.0
1795	19.98%	74.22%	10.0
1816	19.99%	74.24%	10.0
1837	20.01%	74.26%	10.0
1841	20.02%	74.28%	10.0

c) Manufacturer Id vs Loan Default:
https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/ManufacturervsLoanDefault/ManufacturervsLoanDefault-Pareto?publish=yes

As seen in the image below, 81.02% of loan defaults occur in 27.27% of manufacturers.

Man..	% of Loan Default	% of Manufacturer Id	Loan Default
86	44.28%	9.09%	22,410
45	69.84%	18.18%	12,939
51	81.02%	27.27%	5,657
48	90.02%	36.36%	4,554
49	94.44%	45.45%	2,236
120	98.65%	54.55%	2,132
67	99.68%	63.64%	523

d) States vs Loan Defaults
https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/StatesvsLoanDefaults/StatevsLoanDefault-Pareto?publish=yes

As seen in the image below, 81% of loan defaults occur in 40.91% of states.

Stat..	% of Loan Default	% of State ID	Loan Default
4	18.43%	4.55%	9,326
6	32.04%	9.09%	6,890
3	44.58%	13.64%	6,345
13	55.41%	18.18%	5,483
9	62.31%	22.73%	3,492
8	68.75%	27.27%	3,258
14	73.88%	31.82%	2,597
5	77.88%	36.36%	2,023
1	81.00%	40.91%	1,583
11	83.72%	45.45%	1,373
7	86.42%	50.00%	1,369
18	88.78%	54.55%	1,191
2	91.01%	59.09%	1,129
12	93.21%	63.64%	1,118

e) Supplier Id vs Loan Default:
https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/SupplierIdvsLoanDefault/SuppliervsLoanDefault-Pareto?publish=yes

As seen in the image below, 80% of loan defaults occur in 30.10% of supplier Ids.

18110	79.85%	29.94%	15.0
18294	79.88%	29.97%	15.0
18309	79.91%	30.00%	15.0
18312	79.94%	30.04%	15.0
18397	79.97%	30.07%	15.0
18398	80.00%	30.10%	15.0
18415	80.03%	30.14%	15.0
20286	80.06%	30.17%	15.0
21202	80.09%	30.21%	15.0
21475	80.12%	30.24%	15.0

vii. **What are the different employment types given in the data? Can a strategy be developed to fill in the missing values (if any)? Use pie charts to express the different types of employment that define the defaulters and non-defaulters.**

The different Employment types that are given in the data are shown in the below screenshot of the code and the console output:

Code:

```
summary(Loan1$Employment.Type)
```

Output:

```
summary(Loan1$Employment.Type)
   Salaried Self employed          NA's
      97858        127635          7661
```

Here NA's represent the empty cells in the Employment.Type column of the excel data source.

The missing values can be handled in either of the 2 following ways:

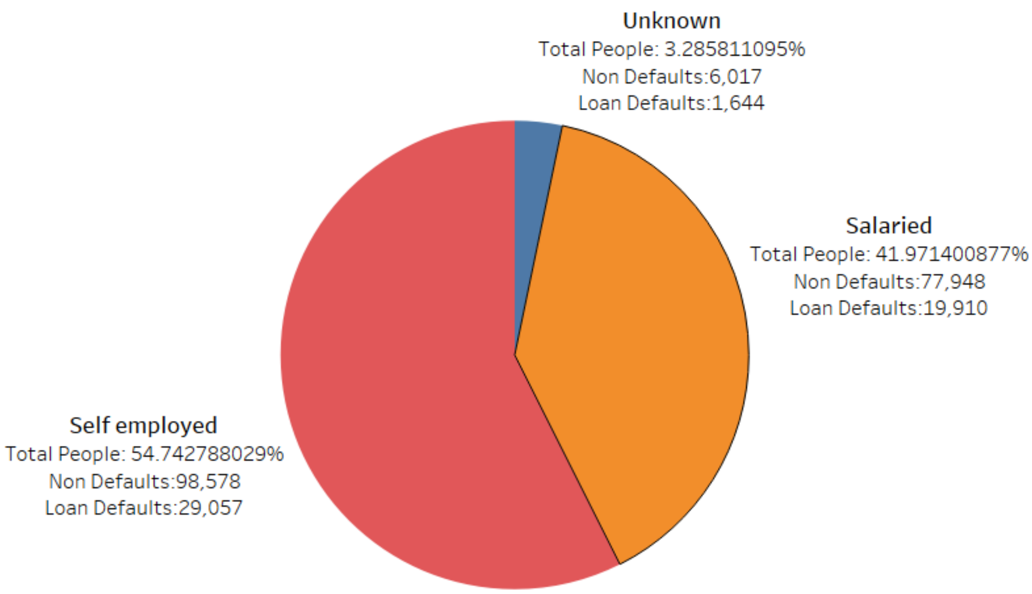
1) Remove the observations where missing values are present.

2) Replace the NA's with "Unknown" and treat it as another level in the factor Employment.Type.

Below is the workbook containing a pie chart to express different types of employments that define the defaulters and non-defaulters.

https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/Piechartfordefaultersineachemploymenttype/Sheet2?publish=yes

Acc. to the pie chart, 54.7% people are self employed with 29057 people out of them being defaulters. Same data inference can be made about other employment types as well.



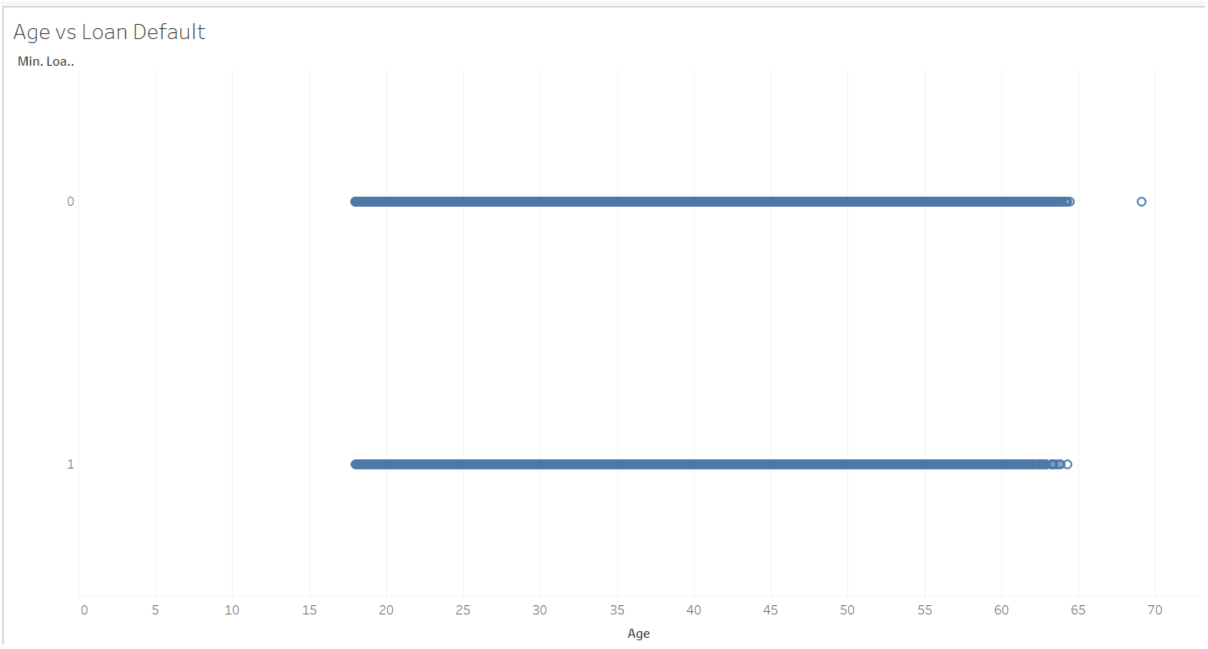
viii. **Has age got anything to do with defaulting? What is the distribution of age w.r.t. to the defaulters and non-defaulters?**

Age can be calculated by using the Date of Birth and Date of Disbursal of loan in the following way:

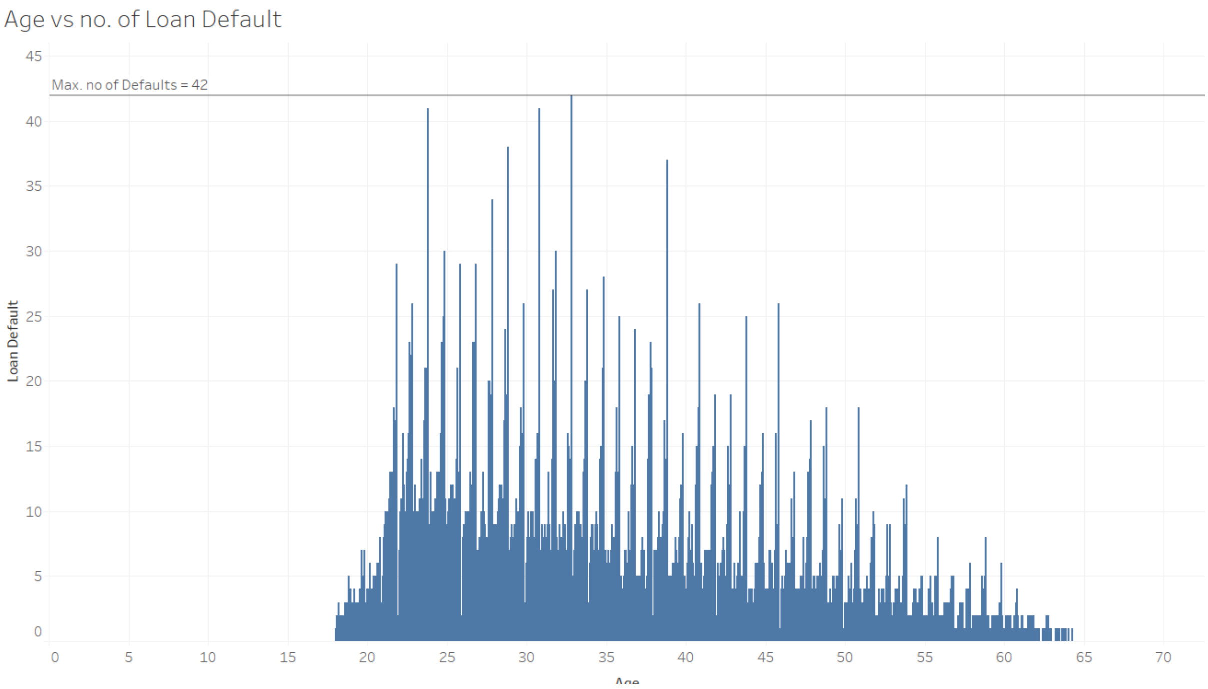
```
#Using date of birth and disbursal date to calculate age at time of disbursal.
#Then removing date of birth and disbursal date as they are not needed anymore.
Loan1$Date.of.Birth = as.Date(Loan1$Date.of.Birth)
Loan1$DisbursalDate = as.Date(Loan1$DisbursalDate)
Loan1$Age = age_calc(Loan1$Date.of.Birth, Loan1$DisbursalDate, units = "years")

Loan1 = Loan1[-c(8,10)]
names(Loan1)
```

https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/AgevsLoanDefault/AgevsLoanDefault?publish=yes



Below is the distribution of number of loan defaults according to Age. A trend is visible is we look in the image below indicating that age might have something to do with defaulting of loans. The significance of Age would be more clear by the model created at the end of this project.



ix. What type of ID was presented by most of the customers for proof?

This can be calculated easily by summing up the variables for each type of ID separately as follows:

Code & Outcome:

```
> sum(Loan$MobileNo_Avl_Flag)
[1] 233154
> sum(Loan$Aadhar_flag)
[1] 195924
> sum(Loan$PAN_flag)
[1] 17621
> sum(Loan$VoterID_flag)
[1] 33794
> sum(Loan$Driving_flag)
[1] 5419
> sum(Loan$Passport_flag)
[1] 496
>
```

According to the above code, Aadhar was given the most times by the customers i.e. 195924 times. Note: Not considering MobileNo_Avl_Flag as an ID. If it is considered so, then it would be the most no. of times shared.

x. Study the credit bureau score distribution. Compare the distribution for defaulters vs. non-defaulters. Explore in detail.

The following table gives the distribution of defaulters and non-defaulters according to various credit bureau score distribution:

https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/CSRdistributionvsLoanDefault/CSRdistributionvsLoanDefaults

CSR distribution vs Loan Defaults

Perform Cns.Score...	Count of train	Loan Default	Non-defaults
No Bureau History A..	116,950	27,052	89,898
C-Very Low Risk	16,045	2,770	13,275
A-Very Low Risk	14,124	2,341	11,783
D-Very Low Risk	11,358	1,699	9,659
B-Very Low Risk	9,201	1,208	7,993
M-Very High Risk	8,776	2,673	6,103
F-Low Risk	8,485	1,580	6,905
K-High Risk	8,277	2,302	5,975
H-Medium Risk	6,855	1,658	5,197
E-Low Risk	5,821	1,000	4,821
I-Medium Risk	5,557	1,515	4,042
G-Low Risk	3,988	786	3,202
Not Scored: Sufficie..	3,765	963	2,802
J-High Risk	3,748	946	2,802
Not Scored: Not Eno..	3,672	770	2,902
Not Scored: No Activ..	2,885	530	2,355
Not Scored: No Upda..	1,534	292	1,242
L-Very High Risk	1,134	318	816
Not Scored: Only a G..	976	208	768
Not Scored: More th..	3	0	3

As seen in the table:
The maximum number of defaults occur in the category where there is "No Beureau History Available" and minimum number of defaults occur where no score is provided due to more than 50 accounts present.

xi. Explore the primary and secondary account details. Is the information in some way related to the loan default probability?

As per the final model created in this document, it can be seen that although Primary account details are factors affecting the loan default probability, Secondary account details seem to be not affecting the probability of loan default.

```
Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)                          -1.401e+00  5.010e-02 -27.954  < 2e-16 ***
disbursed_amount                     -8.978e-02  4.745e-02  -1.892 0.058465 .  
asset_cost                            1.575e-01  3.912e-02   4.028 5.64e-05 ***
ltv                                   4.497e-01  3.911e-02  11.497  < 2e-16 ***
PERFORM_CNS.SCORE                    -8.715e-02  8.544e-03 -10.199  < 2e-16 ***
PRI.ACTIVE.ACCTS                     -2.793e-02  1.441e-02  -1.939 0.052479 .  
PRI.OVERDUE.ACCTS                     1.291e-01  7.545e-03  17.115  < 2e-16 ***
PRI.CURRENT.BALANCE                   6.890e-02  1.656e-02   4.160 3.18e-05 ***
PRI.SANCTIONED.AMOUNT                -3.330e-01  6.286e-02  -5.298 1.17e-07 ***
PRI.DISBURSED.AMOUNT                  2.072e-01  6.400e-02   3.238 0.001206 ** 
NEW.ACCTS.IN.LAST.SIX.MONTHS         -2.823e-02  1.133e-02  -2.491 0.012741 *  
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS   1.011e-01  6.717e-03  15.056  < 2e-16 ***
AVERAGE.ACCT.AGE                      1.228e-01  1.386e-02   8.862  < 2e-16 ***
CREDIT.HISTORY.LENGTH                -1.578e-01  1.632e-02  -9.668  < 2e-16 ***
NO.OF_INQUIRIES                       1.097e-01  6.194e-03  17.706  < 2e-16 ***
Age                                  -8.050e-02  6.824e-03 -11.797  < 2e-16 ***
branch_id                             5.430e-04  9.359e-05   5.802 6.57e-09 ***
supplier_id                           7.645e-06  1.953e-06   3.913 9.10e-05 ***
manufacturer_id                      -3.621e-03  3.016e-04 -12.007  < 2e-16 ***
Current_pincode_ID                    2.750e-05  3.343e-06   8.225  < 2e-16 ***
Employment.TypeSelf employed          1.403e-01  1.339e-02  10.476  < 2e-16 ***
State_ID                              2.247e-02  1.472e-03  15.265  < 2e-16 ***
Employee_code_ID                      3.494e-05  6.622e-06   5.276 1.32e-07 ***
Aadhar_flag                          -2.557e-01  1.908e-02 -13.403  < 2e-16 ***
PAN_flag                             -9.338e-02  2.516e-02  -3.712 0.000206 ***
Driving_flag                         -2.319e-01  4.568e-02  -5.076 3.85e-07 ***
Passport_flag                        -5.130e-01  1.610e-01  -3.186 0.001441 ** 
```

xii. Is there a difference between the sanctioned and disbursed amount of primary and secondary loans? Study the difference by providing appropriate statistics and graphs.

There is a difference in the sanctioned and disbursed amount for primary and secondary accounts resp.

Cumulative difference can be calculated using R as follows:

For Primary:

```
> sum(Loan$PRI.SANCTIONED.AMOUNT) - sum(Loan$PRI.DISBURSED.AMOUNT)
[1] 102111349
> 
```

For Secondary:

```
> sum(Loan$SEC.SANCTIONED.AMOUNT) - sum(Loan$SEC.DISBURSED.AMOUNT)
[1] 27028488
> 
```

Detailed costumer wise difference can be given by the following workbook arranged in descending order:

https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/PrimaryandSecondaryAccountdetails1/PrimaryAccount?publish=yes

xiii. Do customer who make higher number of enquiries end up being higher risk candidates?

This can be found out after making the model. According to the model created, which is shown in the following steps, the following corelation is found in between No. of Queries and loan default probability.

```
> cor(Loan7$NO.OF_INQUIRIES, Loan7$`step0$fitted.values`)
[1] 0.2588955
> 
```

According to the corelation value, it can be seen that the statement “those who make higher number of enquiries end up being higher risk candidates” is not true as no clear pattern can be established between the probability of loan default and no. of queries (due to low corelation value.)

xiv. Is credit history, that is new loans in last six months, loans defaulted in last six months, time since first loan, etc., a significant factor in estimating probability of loan defaulters?

According to the model created as shown in the below image are the significant factors used to identify loan defaults:

As seen, **NEW.ACCTS.IN.LAST.SIX.MONTHS, DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS, CREDIT.HISTORY.LENGTH** are significant factors in estimating the probability of loan defaulters. Although as seem from the p-values for the three, **NEW.ACCTS.IN.LAST.SIX.MONTHS** has highest p-value and is not as significant as the other two factors.

```
Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)                        -1.401e+00  5.010e-02 -27.954  < 2e-16 ***
disbursed_amount                   -8.978e-02  4.745e-02  -1.892 0.058465 .  
asset_cost                          1.575e-01  3.912e-02   4.028 5.64e-05 ***
ltv                                 4.497e-01  3.911e-02  11.497  < 2e-16 ***
PERFORM_CNS.SCORE                  -8.715e-02  8.544e-03 -10.199  < 2e-16 ***
PRI.ACTIVE.ACCTS                   -2.793e-02  1.441e-02  -1.939 0.052479 .  
PRI.OVERDUE.ACCTS                   1.291e-01  7.545e-03  17.115  < 2e-16 ***
PRI.CURRENT.BALANCE                 6.890e-02  1.656e-02   4.160 3.18e-05 ***
PRI.SANCTIONED.AMOUNT              -3.330e-01  6.286e-02  -5.298 1.17e-07 ***
PRI.DISBURSED.AMOUNT                2.072e-01  6.400e-02   3.238 0.001206 ** 
NEW.ACCTS.IN.LAST.SIX.MONTHS       -2.823e-02  1.133e-02  -2.491 0.012741 *  
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS 1.011e-01  6.717e-03  15.056  < 2e-16 ***
AVERAGE.ACCT.AGE                    1.228e-01  1.386e-02   8.862  < 2e-16 ***
CREDIT.HISTORY.LENGTH              -1.578e-01  1.632e-02  -9.668  < 2e-16 ***
NO.OF_INQUIRIES                     1.097e-01  6.194e-03  17.706  < 2e-16 ***
Age                                -8.050e-02  6.824e-03 -11.797  < 2e-16 ***
branch_id                           5.430e-04  9.359e-05   5.802 6.57e-09 ***
supplier_id                         7.645e-06  1.953e-06   3.913 9.10e-05 ***
manufacturer_id                    -3.621e-03  3.016e-04 -12.007  < 2e-16 ***
Current_pincode_ID                  2.750e-05  3.343e-06   8.225  < 2e-16 ***
Employment.TypeSelf employed        1.403e-01  1.339e-02  10.476  < 2e-16 ***
State_ID                            2.247e-02  1.472e-03  15.265  < 2e-16 ***
Employee_code_ID                    3.494e-05  6.622e-06   5.276 1.32e-07 ***
Aadhar_flag                        -2.557e-01  1.908e-02 -13.403  < 2e-16 ***
PAN_flag                           -9.338e-02  2.516e-02  -3.712 0.000206 ***
Driving_flag                       -2.319e-01  4.568e-02  -5.076 3.85e-07 ***
Passport_flag                      -5.130e-01  1.610e-01  -3.186 0.001441 ** 
```

xv. Perform logistic regression modelling, predict the outcome for the test data, and validate the results using the confusion matrix.

```
> #Load the data file
> library(readxl)
> Loan <- read_excel("C:/Users/Vaibhav-PC/Downloads/Project 2/data.xlsx")
>
> #Not performing on orginal df to avoid loading it again and again.
> Loan1 = Loan
>
> #Gives the summary of the variable loan_default
> summary(Loan1$loan_default)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.2171  0.0000  1.0000 
>
> #UniqueID is not required as it is a dummy variable. Thus it can be removed.
> Loan1 = Loan1[-1]
>
> #Converting Employment.Type variable into factor as it is a character vector but should be
> #a categorical variable.
> Loan1$Employment.Type = as.factor(Loan1$Employment.Type)
>
> #To see the distribution of each type of employment.
> summary(Loan1$Employment.Type)
     Salaried Self employed          NA's 
        97858        127635          7661 
>
> #MobileNo_Avl_Flag is affecting no variable & not needed as its min and max are 1.
> #Thus removing it entirely while model creation.
> Loan1 = Loan1[-13]
```

```
> #Converting AVERAGE.ACCT.AGE to numeric values
>
> library(stringr)
> avr = str_split(Loan1$AVERAGE.ACCT.AGE, " ")
> avr1 = 1
> avr2 = 1
> for (i in 1:length(avr)) {avr1[i] = avr[[i]][1]}
> for (i in 1:length(avr)) {avr2[i] = avr[[i]][2]}
> avr1 = gsub("[a-zA-Z]","",avr1)
> avr1 = ifelse(is.na(avr1),0,avr1)
> avr1 = as.numeric(avr1)
>
> avr2 = gsub("[a-zA-Z]","",avr2)
> avr2 = ifelse(is.na(avr2),0,avr2)
> avr2 = as.numeric(avr2)
> avr2 = avr2/12
>
> Loan1$AVERAGE.ACCT.AGE = avr1 + avr2
>
> #Converting CREDIT.HISTORY.LENGTH to numeric values
> avr = str_split(Loan1$CREDIT.HISTORY.LENGTH, " ")
> avr1 = 1
> avr2 = 1
> for (i in 1:length(avr)) {avr1[i] = avr[[i]][1]}
> for (i in 1:length(avr)) {avr2[i] = avr[[i]][2]}
> avr1 = gsub("[a-zA-Z]","",avr1)
> avr1 = ifelse(is.na(avr1),0,avr1)
> avr1 = as.numeric(avr1)
>
> avr2 = gsub("[a-zA-Z]","",avr2)
> avr2 = ifelse(is.na(avr2),0,avr2)
> avr2 = as.numeric(avr2)
> avr2 = avr2/12
>
> Loan1$CREDIT.HISTORY.LENGTH = avr1 + avr2
>
> rm(avr)
> rm(avr1)
> rm(avr2)
> rm(i)
```

```
> #Using date of birth and disbursal date to calculate age at time of disbursal.
> #Then removing date of birth and disbursal date as they are not needed anymore.
> Loan1$Date.of.Birth = as.Date(Loan1$Date.of.Birth)
> Loan1$DisbursalDate = as.Date(Loan1$DisbursalDate)
> library(eeptools)
> Loan1$Age = age_calc(Loan1$Date.of.Birth, Loan1$DisbursalDate, units = "years")
> Loan1 = Loan1[-c(8,10)]
>
> #Outliers removal from disbursed_amount
> LT = mean(Loan1$disbursed_amount) - 2*sd(Loan1$disbursed_amount)
> UT = mean(Loan1$disbursed_amount) + 2*sd(Loan1$disbursed_amount)
>
> Loan2 = subset(Loan1, Loan1$disbursed_amount < UT & Loan1$disbursed_amount > LT)
>
> #Outliers removal from asset_cost
> LT = mean(Loan2$asset_cost) - 2*sd(Loan2$asset_cost)
> UT = mean(Loan2$asset_cost) + 2*sd(Loan2$asset_cost)
>
> Loan3 = subset(Loan2, Loan2$asset_cost < UT & Loan2$asset_cost > LT)
>
> #Making -ve values in PRI.CURRENT.BALANCE as zero.
> Loan3$PRI.CURRENT.BALANCE = ifelse(Loan3$PRI.CURRENT.BALANCE < 0,0,Loan3$PRI.CURRENT.BALANCE)
>
> LT = mean(Loan3$PRI.CURRENT.BALANCE) - 2*sd(Loan3$PRI.CURRENT.BALANCE)
> UT = mean(Loan3$PRI.CURRENT.BALANCE) + 2*sd(Loan3$PRI.CURRENT.BALANCE)
>
> Loan3 = subset(Loan3, Loan3$PRI.CURRENT.BALANCE < UT & Loan3$PRI.CURRENT.BALANCE > LT)
>
```

```
> Loan3 = subset(Loan3, Loan3$PRI.CURRENT.BALANCE < UT & Loan3$PRI.CURRENT.BALANCE > LT)
>
> #Outlier removals in PRI.SANCTIONED.AMOUNT
> LT = mean(Loan3$PRI.SANCTIONED.AMOUNT) - 2*sd(Loan3$PRI.SANCTIONED.AMOUNT)
> UT = mean(Loan3$PRI.SANCTIONED.AMOUNT) + 2*sd(Loan3$PRI.SANCTIONED.AMOUNT)
>
> Loan3 = subset(Loan3, Loan3$PRI.SANCTIONED.AMOUNT < UT & Loan3$PRI.SANCTIONED.AMOUNT > LT)
>
> #Outlier removals in PRI.DISBURSED.AMOUNT
> LT = mean(Loan3$PRI.DISBURSED.AMOUNT) - 2*sd(Loan3$PRI.DISBURSED.AMOUNT)
> UT = mean(Loan3$PRI.DISBURSED.AMOUNT) + 2*sd(Loan3$PRI.DISBURSED.AMOUNT)
>
> Loan3 = subset(Loan3, Loan3$PRI.DISBURSED.AMOUNT < UT & Loan3$PRI.DISBURSED.AMOUNT > LT)
>
> #Removal of -ve values and outliers from SEC.CURRENT.BALANCE
> Loan3$SEC.CURRENT.BALANCE = ifelse(Loan3$SEC.CURRENT.BALANCE < 0, 0, Loan3$SEC.CURRENT.BALANCE)
>
> LT = mean(Loan3$SEC.CURRENT.BALANCE) - 2*sd(Loan3$SEC.CURRENT.BALANCE)
> UT = mean(Loan3$SEC.CURRENT.BALANCE) + 2*sd(Loan3$SEC.CURRENT.BALANCE)
>
> Loan3 = subset(Loan3, Loan3$SEC.CURRENT.BALANCE < UT & Loan3$SEC.CURRENT.BALANCE > LT)
>
> #outlier removal from SEC.SANCTIONED.AMOUNT
> LT = mean(Loan3$SEC.SANCTIONED.AMOUNT) - 2*sd(Loan3$SEC.SANCTIONED.AMOUNT)
> UT = mean(Loan3$SEC.SANCTIONED.AMOUNT) + 2*sd(Loan3$SEC.SANCTIONED.AMOUNT)
>
> Loan3 = subset(Loan3, Loan3$SEC.SANCTIONED.AMOUNT < UT & Loan3$SEC.SANCTIONED.AMOUNT > LT)
>
```

```
> #outlier removal from SEC.DISBURSED.AMOUNT
> LT = mean(Loan3$SEC.DISBURSED.AMOUNT) - 2*sd(Loan3$SEC.DISBURSED.AMOUNT)
> UT = mean(Loan3$SEC.DISBURSED.AMOUNT) + 2*sd(Loan3$SEC.DISBURSED.AMOUNT)
>
> Loan3 = subset(Loan3, Loan3$SEC.DISBURSED.AMOUNT < UT & Loan3$SEC.DISBURSED.AMOUNT > LT)
>
> #outlier removal from PRIMARY.INSTAL.AMT
> LT = mean(Loan3$PRIMARY.INSTAL.AMT) - 2*sd(Loan3$PRIMARY.INSTAL.AMT)
> UT = mean(Loan3$PRIMARY.INSTAL.AMT) + 2*sd(Loan3$PRIMARY.INSTAL.AMT)
>
> Loan3 = subset(Loan3, Loan3$PRIMARY.INSTAL.AMT < UT & Loan3$PRIMARY.INSTAL.AMT > LT)
>
> #outlier removal from SEC.INSTAL.AMT
> LT = mean(Loan3$SEC.INSTAL.AMT) - 2*sd(Loan3$SEC.INSTAL.AMT)
> UT = mean(Loan3$SEC.INSTAL.AMT) + 2*sd(Loan3$SEC.INSTAL.AMT)
>
> Loan3 = subset(Loan3, Loan3$SEC.INSTAL.AMT < UT & Loan3$SEC.INSTAL.AMT > LT)
>
> #Scaling of data frame as it contains numeric variables of with huge variations in range.
> Loan4 = scale(Loan3[c(1,2,3,16,18:36,38)])
> Loan4 = as.data.frame(Loan4)
> Loan4 = cbind(Loan4, Loan3[c(4:15,17,37)])
>
> #Using the approch of omitting the observations with NA's present. This will remove all
> #the observations in Employment.Type that had empty cells in the excel data source.
> Loan5 = na.omit(Loan4)
>
> #Since PERFORM_CNS.SCORE.DESCRIPTION is used to class the score of PERFORM_CNS.SCORE in
> #various categories, thus using the approach of excluding PERFORM_CNS.SCORE.DESCRIPTION
> #in the model creation.
> Loan5 = Loan5[-37]
> #Model Creation
> library(caret)
> set.seed(1)
> intrain = createDataPartition(Loan5$loan_default, p = 0.8, list = F)
> Train = Loan5[intrain,]
> Test = Loan5[-intrain,]
>
> model0 = glm(Train$loan_default ~ ., data = Train, family = binomial(link = "logit"))
> library(MASS)
>
> #Using AIC approach to get the model.
> step0 = stepAIC(model0, direction = "both")
```

After running the stepAIC function, we got the following data model with the least AIC value:

```
Step:  AIC=151817.2
Train$loan_default ~ disbursed_amount + asset_cost + ltv + PERFORM_CNS.SCORE +
    PRI.ACTIVE.ACCTS + PRI.OVERDUE.ACCTS + PRI.CURRENT.BALANCE +
    PRI.SANCTIONED.AMOUNT + PRI.DISBURSED.AMOUNT + NEW.ACCTS.IN.LAST.SIX.MONTHS +
    DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS + AVERAGE.ACCT.AGE +
    CREDIT.HISTORY.LENGTH + NO.OF_INQUIRIES + Age + branch_id +
    supplier_id + manufacturer_id + Current_pincode_ID + Employment.Type +
    State_ID + Employee_code_ID + Aadhar_flag + PAN_flag + Driving_flag +
    Passport_flag

                                      Df Deviance    AIC
<none>                                     151763 151817
+ VoterID_flag                         1   151761 151817
+ PRIMARY.INSTAL.AMT                   1   151762 151818
+ PRI.NO.OF.ACCTS                      1   151762 151818
+ SEC.INSTAL.AMT                       1   151762 151818
- disbursed_amount                     1   151767 151819
+ SEC.OVERDUE.ACCTS                    1   151763 151819
+ SEC.ACTIVE.ACCTS                     1   151763 151819
+ SEC.SANCTIONED.AMOUNT                1   151763 151819
- PRI.ACTIVE.ACCTS                     1   151767 151819
+ SEC.CURRENT.BALANCE                  1   151763 151819
+ SEC.NO.OF.ACCTS                      1   151763 151819
+ SEC.DISBURSED.AMOUNT                 1   151763 151819
- NEW.ACCTS.IN.LAST.SIX.MONTHS         1   151769 151821
- PRI.DISBURSED.AMOUNT                 1   151774 151826
- Passport_flag                        1   151774 151826
- PAN_flag                             1   151777 151829
- supplier_id                          1   151779 151831
- asset_cost                           1   151779 151831
- PRI.CURRENT.BALANCE                  1   151781 151833
- Driving_flag                         1   151790 151842
- Employee_code_ID                     1   151791 151843
- PRI.SANCTIONED.AMOUNT                1   151793 151845
- branch_id                            1   151797 151849
- Current_pincode_ID                   1   151831 151883
- AVERAGE.ACCT.AGE                     1   151842 151894
- CREDIT.HISTORY.LENGTH                1   151863 151915
- PERFORM_CNS.SCORE                    1   151868 151920
- Employment.Type                      1   151873 151925
- ltv                                  1   151894 151946
- Age                                  1   151904 151956
- manufacturer_id                      1   151908 151960
- Aadhar_flag                          1   151940 151992
- DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS  1   151984 152036
- State_ID                             1   151994 152046
- PRI.OVERDUE.ACCTS                    1   152055 152107
- NO.OF_INQUIRIES                      1   152077 152129
> summary(step0)

Call:
glm(formula = Train$loan_default ~ disbursed_amount + asset_cost +
    ltv + PERFORM_CNS.SCORE + PRI.ACTIVE.ACCTS + PRI.OVERDUE.ACCTS +
    PRI.CURRENT.BALANCE + PRI.SANCTIONED.AMOUNT + PRI.DISBURSED.AMOUNT +
    NEW.ACCTS.IN.LAST.SIX.MONTHS + DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS +
    AVERAGE.ACCT.AGE + CREDIT.HISTORY.LENGTH + NO.OF_INQUIRIES +
    Age + branch_id + supplier_id + manufacturer_id + Current_pincode_ID +
    Employment.Type + State_ID + Employee_code_ID + Aadhar_flag +
    PAN_flag + Driving_flag + Passport_flag, family = binomial(link = "logit"),
    data = Train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1843  -0.7538  -0.6410  -0.4271   2.6810

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -1.401e+00  5.010e-02 -27.954  < 2e-16 ***
disbursed_amount                    -8.978e-02  4.745e-02  -1.892 0.058465 .
asset_cost                           1.575e-01  3.912e-02   4.028 5.64e-05 ***
ltv                                  4.497e-01  3.911e-02  11.497  < 2e-16 ***
PERFORM_CNS.SCORE                   -8.715e-02  8.544e-03 -10.199  < 2e-16 ***
PRI.ACTIVE.ACCTS                    -2.793e-02  1.441e-02  -1.939 0.052479 .
PRI.OVERDUE.ACCTS                    1.291e-01  7.545e-03  17.115  < 2e-16 ***
PRI.CURRENT.BALANCE                  6.890e-02  1.656e-02   4.160 3.18e-05 ***
PRI.SANCTIONED.AMOUNT               -3.330e-01  6.286e-02  -5.298 1.17e-07 ***
PRI.DISBURSED.AMOUNT                 2.072e-01  6.400e-02   3.238 0.001206 **
NEW.ACCTS.IN.LAST.SIX.MONTHS        -2.823e-02  1.133e-02  -2.491 0.012741 *
DELINQUENT.ACCTS.IN.LAST.SIX.MONTHS  1.011e-01  6.717e-03  15.056  < 2e-16 ***
AVERAGE.ACCT.AGE                     1.228e-01  1.386e-02   8.862  < 2e-16 ***
CREDIT.HISTORY.LENGTH               -1.578e-01  1.632e-02  -9.668  < 2e-16 ***
NO.OF_INQUIRIES                      1.097e-01  6.194e-03  17.706  < 2e-16 ***
Age                                 -8.050e-02  6.824e-03 -11.797  < 2e-16 ***
branch_id                            5.430e-04  9.359e-05   5.802 6.57e-09 ***
supplier_id                          7.645e-06  1.953e-06   3.913 9.10e-05 ***
manufacturer_id                     -3.621e-03  3.016e-04 -12.007  < 2e-16 ***
Current_pincode_ID                   2.750e-05  3.343e-06   8.225  < 2e-16 ***
Employment.TypeSelf employed         1.403e-01  1.339e-02  10.476  < 2e-16 ***
State_ID                             2.247e-02  1.472e-03  15.265  < 2e-16 ***
Employee_code_ID                     3.494e-05  6.622e-06   5.276 1.32e-07 ***
Aadhar_flag                         -2.557e-01  1.908e-02 -13.403  < 2e-16 ***
PAN_flag                            -9.338e-02  2.516e-02  -3.712 0.000206 ***
Driving_flag                        -2.319e-01  4.568e-02  -5.076 3.85e-07 ***
Passport_flag                       -5.130e-01  1.610e-01  -3.186 0.001441 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157104  on 147951  degrees of freedom
Residual deviance: 151763  on 147925  degrees of freedom
AIC: 151817

Number of Fisher Scoring iterations: 4

>
```

The final confusion matrix that was given by this model is as follows:

```
> #Predicting values using model in the Test data created using createDataPartition function.
> Pred = predict(step0, newdata = Test[,-37], type = "response")
> Pred1 = ifelse(Pred < 0.4, 0, 1)
> #Create a confusion matrix
> library(e1071)
> a = table(Test$loan_default, Pred1, dnn = list("actual", "predicted"))
> a
      predicted
actual     0     1
     0 28265   424
     1  7981   317
> caret::confusionMatrix(a)
Confusion Matrix and Statistics

      predicted
actual     0     1
     0 28265   424
     1  7981   317

               Accuracy : 0.7728
                 95% CI : (0.7685, 0.777)
    No Information Rate : 0.98
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0346

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7798
            Specificity : 0.4278
         Pos Pred Value : 0.9852
         Neg Pred Value : 0.0382
             Prevalence : 0.9800
         Detection Rate : 0.7642
   Detection Prevalence : 0.7757
      Balanced Accuracy : 0.6038

       'Positive' Class : 0

>
```

Similarly other models can be created by either changing the value of "p" in createDataPartition function and/or selecting a different threshold value to define 0 and 1 in the **Pred1** object vector created for the confusion matrix.

The model with the highest Accuracy value in the confusion matrix is preferred.

Dashboarding:

xvi. Visualize the data using Tableau to help user explore data to have a better understanding

Created a story to explain the loan defaults according to various variables. This story can even be expanded. But just for the sake of explaining, used half of the independent variables of the total found in the final model above.

https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/FinalProject2Workbook/Story

xvii. Demonstrate the variables associated with each other and factors to build a dashboard

Created a dashboard showing the relationship between variables like CNS Score description, Employment type, Age, State ID to show the number of loan defaults. Made then as filters to make them change if any one of the field is highlighted in any sheet of the dashboard.

https://public.tableau.com/profile/vaibhav.bajaj#!/vizhome/FinalProject2Workbook/Dashboard