

Three blue spheres of varying sizes are positioned in the upper right and lower right areas of the page. Thin blue lines extend from the top left and top right corners towards the spheres.

## Bank Loan Default Prediction Model

Mini Project

Build a model to predict default loan that will help the bank to take required actions.

**NIVEDITA DEY - PGP BABI May'19**  
**29/03/2020**

## Contents

1. Project Introduction.....	2
2. Data Report.....	3
3. Exploratory Data Analysis .....	4
3.1 Univariate Analysis.....	4
3.2 Bi-Variate Analysis.....	5
3.3 Unwanted Variable .....	5
3.4 Missing Value Identification.....	6
3.5 Outlier Identification.....	6
3.6 Variable Transformation / Feature Creation .....	6
3.7 Correlation/Multicollinearity .....	7
3.8 Addition of New Variable.....	7
4. Insights from EDA.....	8

## 1. Project Introduction

### A) Defining problem statement

Retail lending is generally considered risk-free as the bank might have done necessary due diligence including collateral requirement and credit score. However, it has been recently witnessed that this segment has started to default, in turn impacting the revenue and profitability for the bank.

Hence, there is a need to build a model to predict default loan which will help the bank to take required actions including –

- i. Avoiding the exposure
- ii. Intensify the collection efforts
- iii. Initiate the collateral sale
- iv. Avoiding certain customer or product segment

### B) Need of the study/project

Retail lending is an important division of any large bank and it helps the bank to grow at rapid pace by earning significant fee income, interest income while also fetching savings and current account.

In some of the banks, it contributes more than 70% of the banks' assets and revenues. In some cases, higher delinquency has even led to bank run and closure of a bank. Recent example of YES Bank is well known where lending to risky customers led to unprecedented by RBI and impacted the brand image of the bank.

### C) Understanding business/social opportunity

If the bank is able to effectively predict the chance of loan default before the disbursement, then the future delinquency can be reduced significantly. This will help the bank to maintain good profitability and avoid any capital erosion.

This model will also enable the bank to identify the customer and product segments which have lower delinquency and high profitability. It will help the bank to expand into new territories and segments where they have not ventured before (for e.g., tier III, tier IV towns).

## 2. Data Report

We attempt to predict the risk of the loan being default based on the past loan data. Hence, we will take an overview the given data:

### A) Understanding how data was collected in terms of time, frequency and methodology

The data contains the details of Loans which have been issued between June 2007 and December 2015 period.

Maximum last payment date for the loans is January 2016. Hence we can consider data is collected post January 2016. Based on the loan issue date, it shows Monthly frequency of data collection.

### B) Visual inspection of data (rows, columns, descriptive details)

There are 226,786 rows and 41 columns.

Out of which 25 are numeric columns, 11 character columns and 5 date columns.

The last variable 'loan\_status' is the dependent variable.

### C) Understanding of attributes (variable info, renaming if required)

We have renamed the column 'earliest\_cr\_line' to 'earliest\_cr\_line\_mnth' as it shows the month a borrower's earliest reported credit line was opened.

### 3. Exploratory Data Analysis

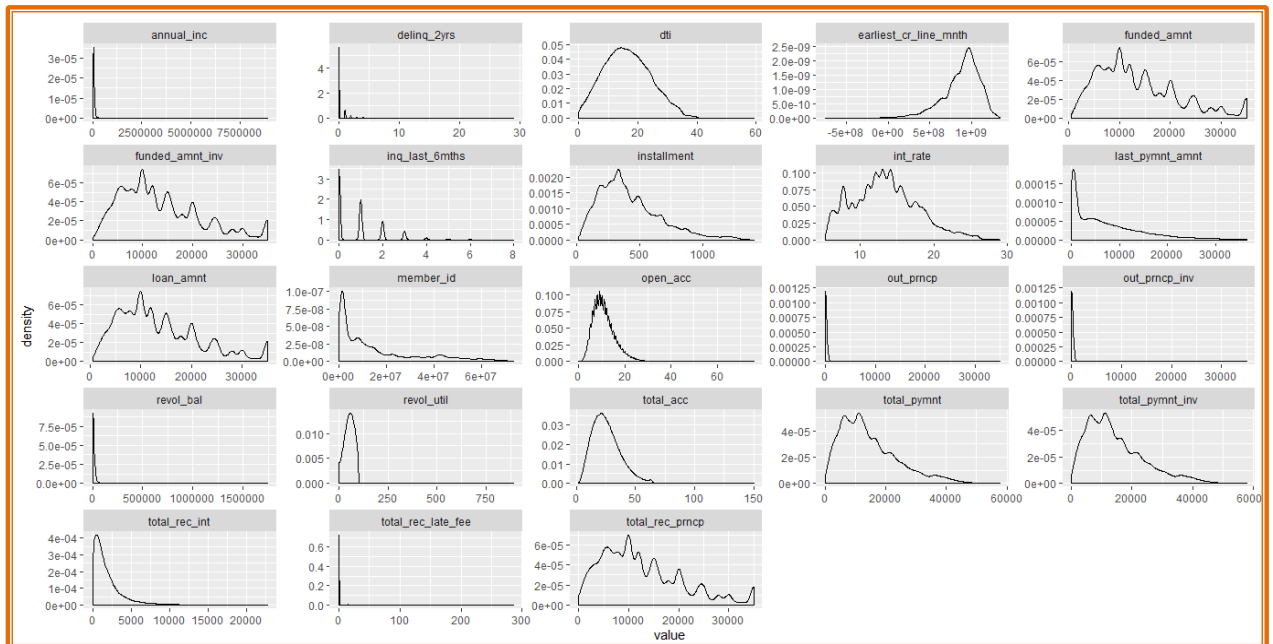
The various steps followed to analyze the case study is mentioned and explained below.

#### 3.1 Univariate Analysis

We are analyzing the all the 41 independent variable from data set give which we have stored in the data frame 'loanData'. The 'loan\_status' variable is the dependent variable.

We perform Univariate analysis.

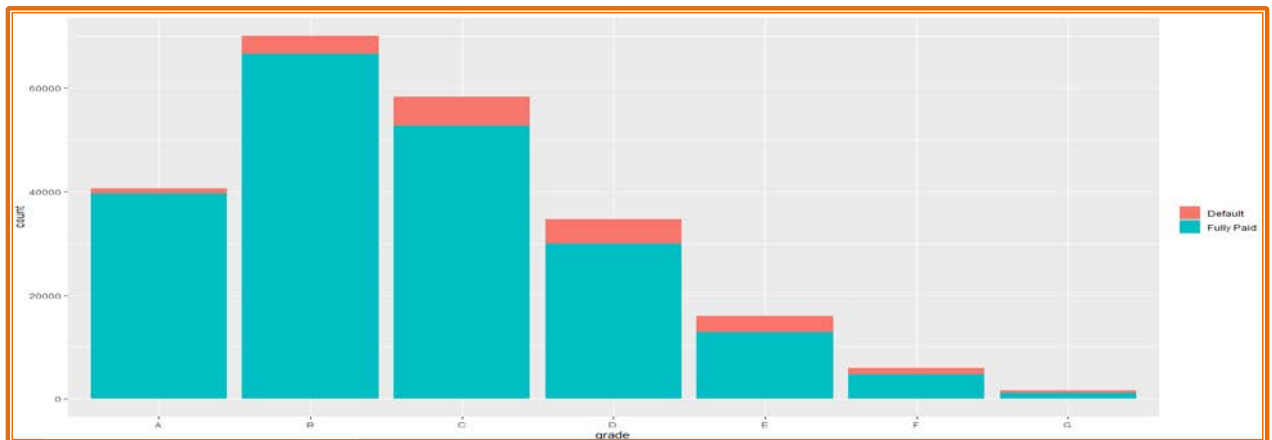
- Nearly 80% customers have 0 number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years which confirms that these customers have good track record
- Most of the loans are disbursed in the range of 5000 to 15000.
- Around 80% of customer who have availed loan has annual income less than 100,000
- Large part of customer has been lent within 30% of DTI. This gives comfort about the loan portfolio
- **Few variables like revol\_bal, out\_prncp, out\_prncp\_inv and total\_rec\_late\_fee** are concentrated to a particular range of values. Hence there is difference in mean and median.
- The summary and box plot shows there is an **outlier** in most of the continuous variables. On further analysis we found that those are acceptable values.



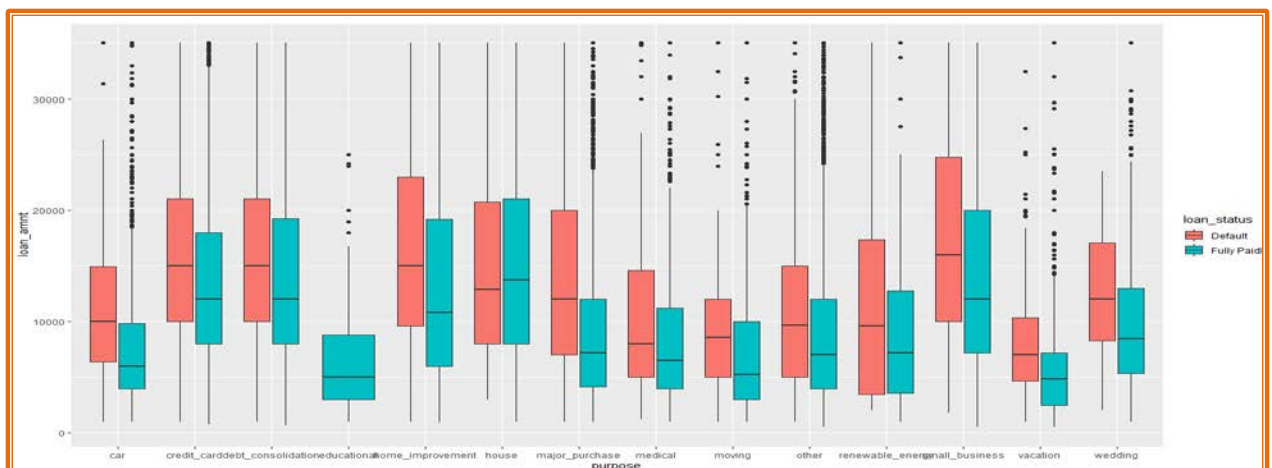
Please refer Appendix A for Source Code.

### 3.2 Bi-Variate Analysis

We will analyze loan\_status with the independent variables from data set 'loanData'. Most of the variables do not seem to have much effect whether loan will default or not. Customers who have mostly defaulted belong to E, F and G grade compared to fully paid in the same grade. Customers with loan Grade B have fully paid the loan maximum time.



Customers have borrowed higher amount of loan mostly for credit card, debt consolidation, home improvement, house and small business



Please refer Appendix A for Source Code

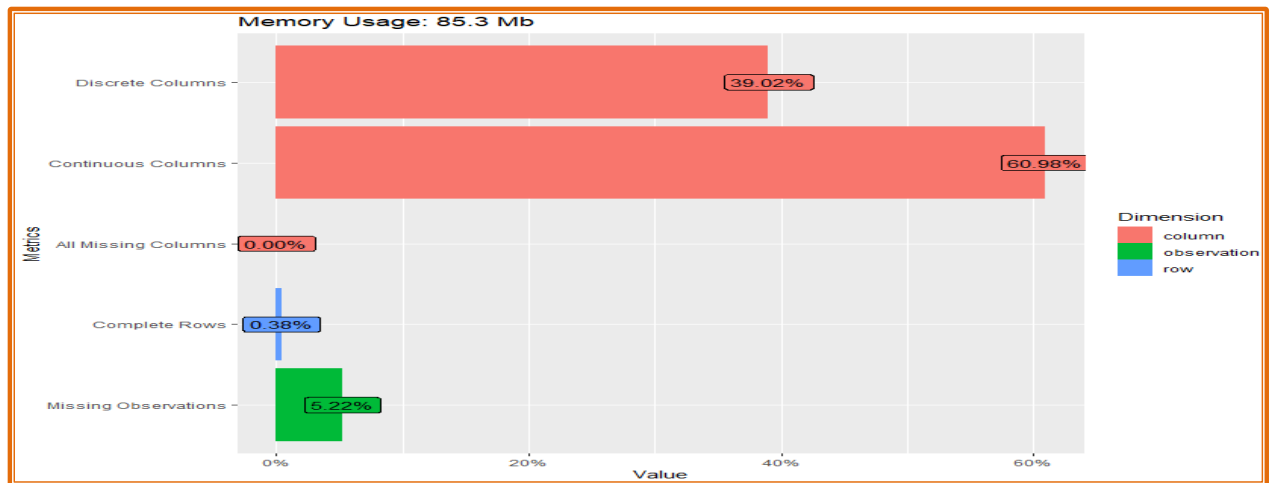
### 3.3 Unwanted Variable

The list of unwanted variables whichj we are not using in our analysis is given below –

Variable	Reason
recoveries	All values are 0
collection_recovery_fee	All values are 0
pymnt_plan	Only 6 rows (4-Default and 2-Fully paid) with value as y
desc	It is detail briefing on 'Purpose' variable. Hence redundant information
mths_since_last_delinq	Nearly 54% missing value so removed it. Explained in Missing value section

### 3.4 Missing Value Identification

We use 'is na' function to check if there are any missing values. There are missing values. Hence we plot to get the overall status



Variable	Missing Count	Treatment
next_pymnt_d	207723	Verified that the values are missing where loans are Fully paid. Hence it is valid scenario
mths_since_last_delinq	124638	has nearly 54% of missing data. Hence we will remove the variable from our analysis
revol_util	164	Imputed with mean
last_pymnt_d	341	Verified that since inception there is no repayment of the loan Hence valid scenario
last_credit_pull_d	16	Removed the rows
Desc	152077	From our study the data we understand that it is the detail of the 'Purpose' variable. Hence we drop desc as it will be redundant information

Please refer Appendix A for Source Code.

### 3.5 Outlier Identification

There are outliers in very most of the variables. It is evident from the box plot and summary as well but the values are acceptable. Hence we did not treat the outliers

Please refer Appendix A for Source Code.

### 3.6 Variable Transformation / Feature Creation

In Summary of data we have seen that few of the variables are character. So we use 'as.factor' to convert them.

We also convert the Date variables to number for ease in analysis.

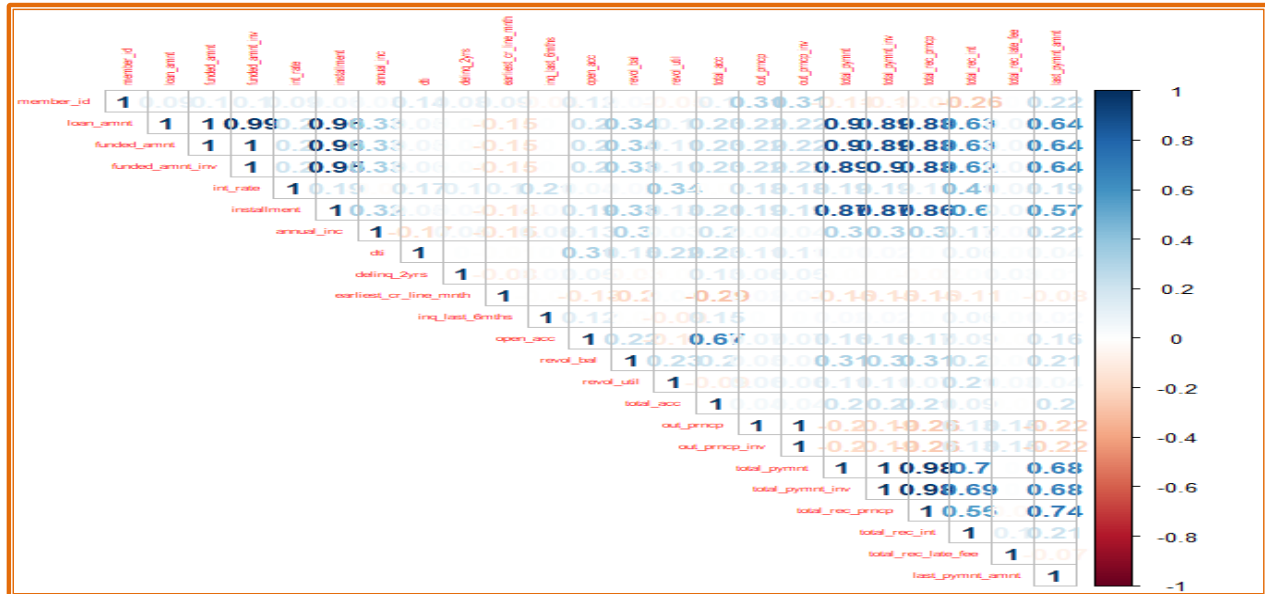
The Dependent variable 'loan\_status' is converted in terms of 0 and 1

Fully Paid: 0

Default: 1

Please refer Appendix A for Source Code.

### 3.7 Correlation/Multicollinearity



Based on the above plot we can see few of the variables are highly co-related. Hence we find the list of highly correlated independent variables.

```
> high_corr <- findCorrelation(loanCor, cutoff = .8)
> high_corr = getNumericColumns(loanDataNum)[high_corr]
> high_corr
[1] "funded_amnt"      "loan_amnt"        "funded_amnt_inv"  "total_pymnt"      "total_pymnt_inv"
[6] "installment"     "out_prncp"
```

For further calibration we check the VIF value (multicollinearity) as well.

For the above variables, the VIF is more than 10 and **hence we drop them**

Please refer Appendix A for Source Code.

### 3.8 Addition of New Variable

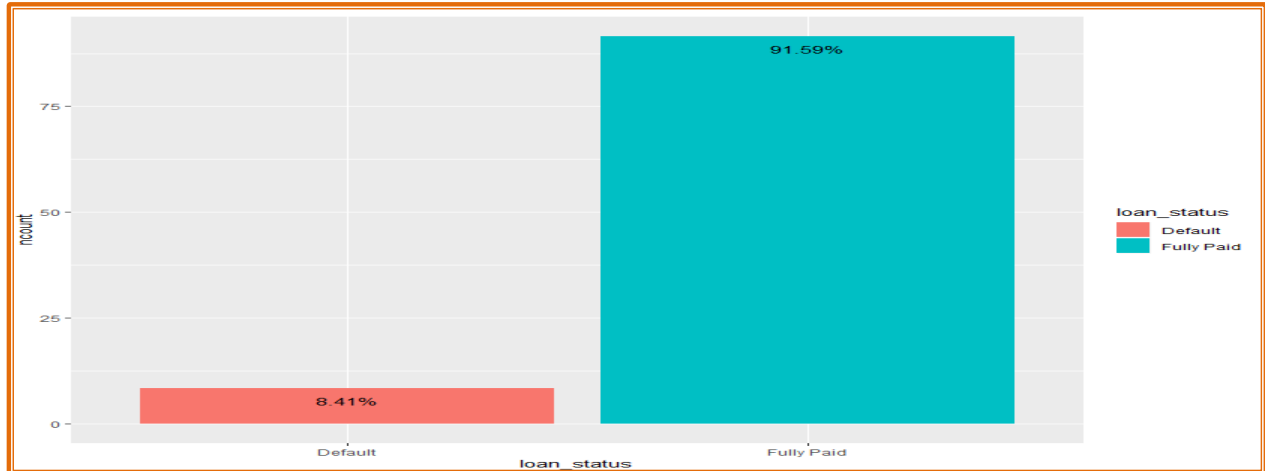
All the key variables required for building the model like Debt to Income ratio or Fixed obligations to income ratio, revolving balance utilization rate, outstanding principal, past delinquency etc. are available in the given data set.

Important variable not available in the dataset is the collateral value which bank might have taken for the loan. If the collateral is available at >100% of the loan amount, it gives comfort to the bank to give specific loans.

Hence, no new variable is required to be created.

## 4. Insights from EDA

The data is highly imbalanced. So while building the model we can either choose to undersample the minority class or oversample the majority class (SMOTE)



**Significant Variable:** The list of highly significant variables is: member\_id, terms, installment, grade, emp\_length, dti, issue\_d, revol\_bal, revol\_util, total\_rec\_int, total\_rec\_late\_fee, last\_pymnt\_d, last\_pymnt\_amnt, last\_credit\_pull\_d