Problem Statement: Predict the interest rate on the loan given pertaining parameters related to loan. Build machine learning/statistical models in R to predict the interest rate assigned to a loan.

Names of Attributes

There are a total of 32 attributes named from X1 to X32 with 4 lac observations where

X1	Interest Rate on the loan
X2	A unique id for the loan.
X3	A unique id assigned for the borrower.
X4	Loan amount requested
X5	Loan amount funded
X6	Investor-funded portion of loan
X7	Number of payments (36 or 60)
X8	Loan grade
X9	Loan subgrade
X10	Employer or job title (self-filled)
X11	Number of years employed (0 to 10; 10 = 10 or more)
X12	Home ownership status: RENT, OWN, MORTGAGE, OTHER.
X13	Annual income of borrower
X14	Income verified, not verified, or income source was verified
X15	Date loan was issued
X16	Reason for loan provided by borrower
X17	Loan category, as provided by borrower
X18	Loan title, as provided by borrower
X19	First 3 numbers of zip code
X20	State of borrower
X21	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's self-reported monthly income.
X22	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
X23	Date the borrower's earliest reported credit line was opened
X24	Number of inquiries by creditors during the past 6 months.
X25	Number of months since the borrower's last delinquency.
X26	Number of months since the last public record.
X27	Number of open credit lines in the borrower's credit file.
X28	Number of derogatory public records
X29	Total credit revolving balance
X30	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
X31	The total number of credit lines currently in the borrower's credit file
X32	The initial listing status of the loan. Possible values are W, F

Structure of Data

After fixing attributes to their respective types and replacing empty values with NA's

```
> str(data)
'data.frame':   400000 obs. of  32 variables:
 $ X1 : num  11.9 10.7 17 13.1 13.6 ...
 $ X2 : num  54734 55742 57167 57245 57416 ...
 $ X3 : num  80364 114426 137225 138150 139635 ...
 $ X4 : num  25000 7000 25000 1200 10800 7200 7500 3000 4000 5600 ...
 $ X5 : num  25000 7000 25000 1200 10800 ...
 $ X6 : num  19080 673 24725 1200 10692 ...
 $ X7 : Factor w/ 2 levels " 36 months"," 60 months": 1 1 1 1 1 1 1 1 1 1 ...
 $ X8 : Factor w/ 7 levels "A","B","C","D",..: 2 2 4 3 3 4 2 3 1 4 ...
 $ X9 : Factor w/ 35 levels "A1","A2","A3",..: 9 10 18 12 13 19 8 15 5 17 ...
 $ X10: chr  NA "CNN" "Web Programmer" "city of beaumont texas" ...
 $ X11: Factor w/ 12 levels "< 1 year","1 year",..: 1 1 2 3 8 11 5 5 1 2 ...
 $ X12: Factor w/ 6 levels "ANY","MORTGAGE",..: 6 6 6 5 6 6 6 2 2 6 ...
 $ X13: num  85000 65000 70000 54000 32000 58000 85000 80800 148000 45000 ...
 $ X14: Factor w/ 3 levels "not verified",..: 2 1 2 1 1 3 1 1 1 1 ...
 $ X15: Factor w/ 91 levels "10-Apr","10-Aug",..: 81 76 50 8 89 26 68 91 4 4 ...
 $ X16: chr  "Due to a lack of personal finance education and exposure to poor financing skills growing up, I was easy prey
ant to pay off the last bit of credit card debt at a better rate." "Trying to pay a friend back for apartment broker's fee
." "If funded, I would use this loan consolidate two loans with interest rates of 15 and 16 percent respectively.   I have
 $ X17: Factor w/ 14 levels "car","credit_card",..: 3 2 3 3 3 3 3 2 2 3 ...
 $ X18: chr  "Debt consolidation for on-time payer" "Credit Card payoff" "mlue" "zxcvb" ...
 $ X19: Factor w/ 877 levels "007xx","008xx",..: 822 95 83 682 53 21 803 680 680 187 ...
 $ X20: Factor w/ 50 levels "AK","AL","AR",..: 5 34 34 43 7 39 5 43 43 21 ...
 $ X21: num  1948 1429 1050 547 1163 ...
 $ X22: num  0 0 0 0 0 0 0 1 0 0 ...
 $ X23: chr  "Feb-94" "Oct-00" "Jun-00" "Jan-85" ...
 $ X24: num  0 0 0 0 1 0 1 0 0 0 ...
 $ X25: num  NA NA 41 64 58 26 NA 13 NA 38 ...
 $ X26: num  NA NA NA NA NA NA NA 0 NA 63 ...
 $ X27: num  10 7 10 5 14 6 3 13 11 5 ...
 $ X28: num  0 0 0 0 0 0 0 0 0 1 ...
 $ X29: num  28854 33623 19878 2584 3511 ...
 $ X30: num  52.1 76.7 66.3 40.4 25.6 90.1 73.2 39.5 51 76.8 ...
 $ X31: num  42 7 17 31 40 25 11 23 19 9 ...
 $ X32: Factor w/ 2 levels "f","w": 1 1 1 1 1 1 1 1 1 1 ...
```

Featuring Engineering

One variable is created as following:

credit_limit – which is

$$\frac{\boldsymbol{total\ credit\ revolving\ balance\ (X29)}}{\boldsymbol{Revolving\ line\ utilization\ rate\ (X30)}} * \mathbf{100}$$

Summary of Data

```
> summary(data)
       X1                X2                X3                X4                X5                X6                X7                X8                X9
 Min.   : 5.42   Min.   :   54734   Min.   :   70699   Min.   :  500   Min.   :  500   Min.   :    0   36 months:292369   B      :101668   B3     : 24009
 1st Qu.:10.99   1st Qu.: 3151742   1st Qu.: 3727712   1st Qu.: 8000   1st Qu.: 8000   1st Qu.: 8000   60 months:107630   C      : 90071   B4     : 22611
 Median :13.68   Median : 8234778   Median : 9667699   Median :12000   Median :12000   Median :12000   NA's      :     1   D      : 55621   B2     : 19853
 Mean   :13.95   Mean   : 9984493   Mean   :11338986   Mean   :14274   Mean   :14246   Mean   :14183                       A      : 53707   C1     : 19285
 3rd Qu.:16.78   3rd Qu.:15329598   3rd Qu.:17312192   3rd Qu.:20000   3rd Qu.:20000   3rd Qu.:19900                       E      : 25518   C2     : 19182
 Max.   :26.06   Max.   :28753146   Max.   :31278050   Max.   :35000   Max.   :35000   Max.   :35000                       (Other): 12145   (Other):233790
 NA's   :61010   NA's   :1          NA's   :1          NA's   :1       NA's   :1       NA's   :1                           NA's   : 61270   NA's   : 61270
     X10                X11               X12                X13                                  X14                X15             X16
 Length:400000      10+ years:128060   ANY     :     1   Min.   :   3000   not verified            :127220   14-Jul : 29306   Length:400000
 Class :character   2 years  : 35427   MORTGAGE:172112   1st Qu.:  45000   VERIFIED - income       :149686   14-May : 19099   Class :character
 Mode  :character   3 years  : 31428   NONE    :    36   Median :  63000   VERIFIED - income source:123093   14-Apr : 19071   Mode  :character
                    < 1 year : 30607   OTHER   :   124   Mean   :  73160   NA's                    :     1   14-Aug : 18814
                    5 years  : 27277   OWN     : 29588   3rd Qu.:  88200                                      14-Jun : 17179
                    (Other)  :147200   RENT    :136778   Max.   :7500000                                      (Other):296530
                    NA's     :     1   NA's    : 61361   NA's   :61028                                        NA's   :     1
                X17             X18                X19               X20              X21             X22               X23                X24
 debt_consolidation:233794   Length:400000      945xx  :  4622   CA     : 62194   Min.   :   1   Min.   : 0.0000   Length:400000      Min.   :0.0000
 credit_card       : 89484   Class :character   112xx  :  4391   NY     : 34557   1st Qu.:1125   1st Qu.: 0.0000   Class :character   1st Qu.:0.0000
 home_improvement  : 23140   Mode  :character   750xx  :  4341   TX     : 31277   Median :1670   Median : 0.0000   Mode  :character   Median :0.0000
 other             : 20161                      606xx  :  4041   FL     : 26991   Mean   :1700   Mean   : 0.2745                      Mean   :0.8172
 major_purchase    :  8664                      100xx  :  3834   IL     : 15877   3rd Qu.:2252   3rd Qu.: 0.0000                      3rd Qu.:1.0000
 (Other)           : 24756                      (Other):378770   (Other):229103   Max.   :3955   Max.   :29.0000                      Max.   :8.0000
 NA's              :     1                      NA's   :     1   NA's   :     1   NA's   :1      NA's   :1                            NA's   :1
      X25               X26               X27               X28               X29               X30               X31           X32
 Min.   :  0.00   Min.   :  0.0   Min.   : 0.00   Min.   : 0.0000   Min.   :     0   Min.   :  0.00   Min.   :  2.00   f   :274313
 1st Qu.: 16.00   1st Qu.: 54.0   1st Qu.: 8.00   1st Qu.: 0.0000   1st Qu.:  6453   1st Qu.: 39.50   1st Qu.: 17.00   w   :125686
 Median : 31.00   Median : 80.0   Median :10.00   Median : 0.0000   Median :  11778   Median : 57.80   Median : 23.00   NA's:     1
 Mean   : 34.31   Mean   : 76.3   Mean   :11.12   Mean   : 0.1523   Mean   :  15956   Mean   : 56.28   Mean   : 24.98
 3rd Qu.: 50.00   3rd Qu.:103.0   3rd Qu.:14.00   3rd Qu.: 0.0000   3rd Qu.:  20209   3rd Qu.: 74.90   3rd Qu.: 32.00
 Max.   :188.00   Max.   :129.0   Max.   :76.00   Max.   :63.0000   Max.   :2568995   Max.   :892.30   Max.   :121.00
 NA's   :218802   NA's   :348845  NA's   :1       NA's   :1         NA's   :1         NA's   :267      NA's   :1
```

NA's & Outliers are removed from some attributes which we need for further processing.

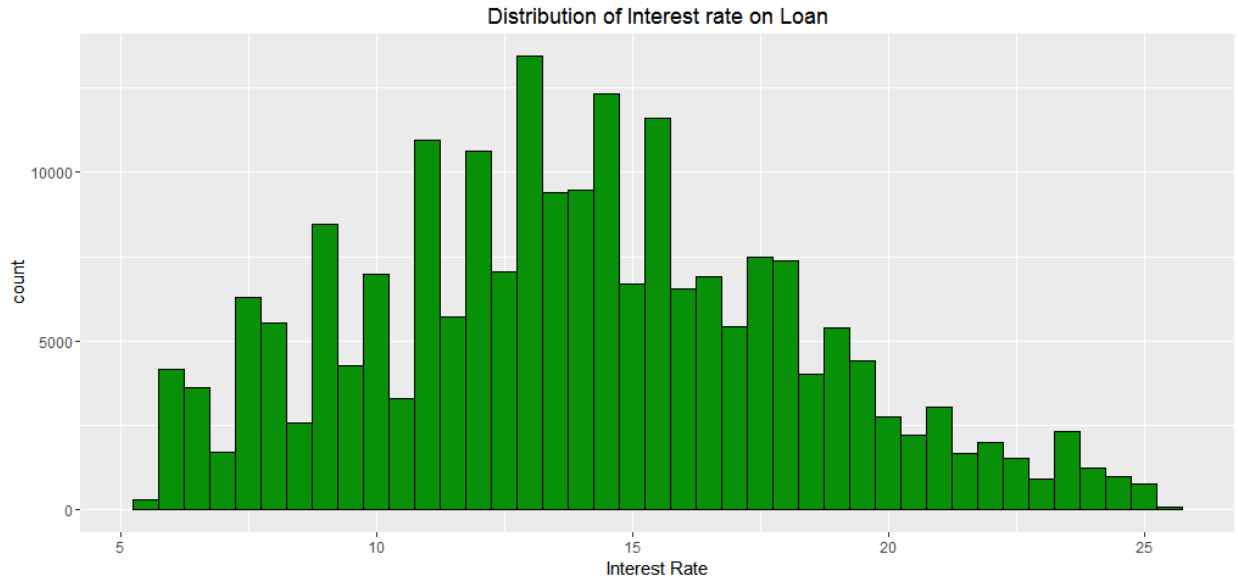
Correlation b/w attributes is as follows:

```
> corr
                   X1      X4      X5      X6      X13     X21     X22     X24     X27     X28     X29     X30     X31 credit_limit
X1            1.0000  0.1910  0.1920  0.1949 -0.0330  0.1536  0.0860  0.2100  0.0408  0.0642  0.0705  0.3277 -0.0192      -0.1812
X4            0.1910  1.0000  0.9984  0.9940  0.4572  0.0740  0.0105 -0.0166  0.1674 -0.0711  0.4729  0.1408  0.2146       0.4016
X5            0.1920  0.9984  1.0000  0.9959  0.4568  0.0754  0.0111 -0.0170  0.1684 -0.0707  0.4731  0.1419  0.2147       0.4008
X6            0.1949  0.9940  0.9959  1.0000  0.4552  0.0798  0.0119 -0.0186  0.1702 -0.0690  0.4719  0.1439  0.2161       0.3980
X13          -0.0330  0.4572  0.4568  0.4552  1.0000 -0.2010  0.0907  0.0693  0.1771 -0.0191  0.3373  0.0772  0.2886       0.3056
X21           0.1536  0.0740  0.0754  0.0798 -0.2010  1.0000  0.0014  0.0062  0.3184 -0.0415  0.2597  0.1917  0.2439       0.1359
X22           0.0860  0.0105  0.0111  0.0119  0.0907  0.0014  1.0000  0.0263  0.0680 -0.0141 -0.0597 -0.0170  0.1468      -0.0598
X24           0.2100 -0.0166 -0.0170 -0.0186  0.0693  0.0062  0.0263  1.0000  0.0931  0.0369 -0.0607 -0.0995  0.1237      -0.0037
X27           0.0408  0.1674  0.1684  0.1702  0.1771  0.3184  0.0680  0.0931  1.0000 -0.0167  0.2925 -0.1066  0.6360       0.4102
X28           0.0642 -0.0711 -0.0707 -0.0690 -0.0191 -0.0415 -0.0141  0.0369 -0.0167  1.0000 -0.1482 -0.0703  0.0194      -0.1332
X29           0.0705  0.4729  0.4731  0.4719  0.3373  0.2597 -0.0597 -0.0607  0.2925 -0.1482  1.0000  0.4312  0.2486       0.7246
X30           0.3277  0.1408  0.1419  0.1439  0.0772  0.1917 -0.0170 -0.0995 -0.1066 -0.0703  0.4312  1.0000 -0.0727      -0.1951
X31          -0.0192  0.2146  0.2147  0.2161  0.2886  0.2439  0.1468  0.1237  0.6360  0.0194  0.2486 -0.0727  1.0000       0.3265
credit_limit -0.1812  0.4016  0.4008  0.3980  0.3056  0.1359 -0.0598 -0.0037  0.4102 -0.1332  0.7246 -0.1951  0.3265       1.0000
```

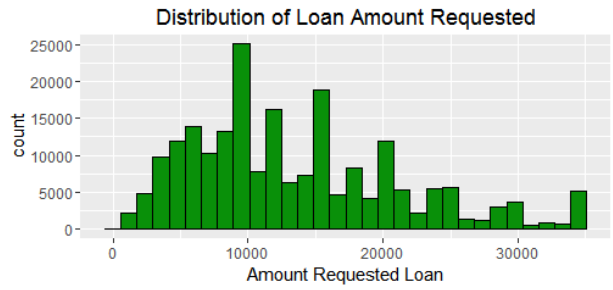
From here we can see that attributes X4, X5 & X6 are highly collinear which can create problem of multi-co linearity when fed to a model.

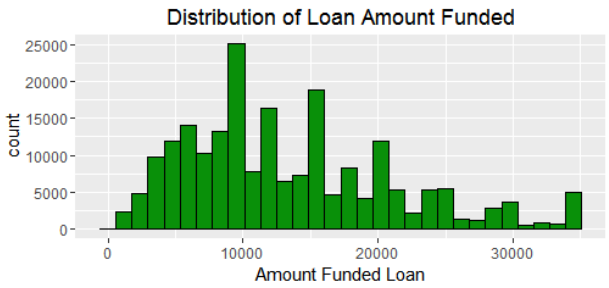
Rest of the attributes doesn't show correlation b/w them as much.

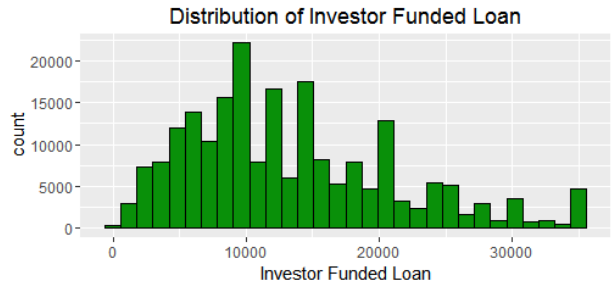
Let’s see the distribution of Interest Rate (X1)



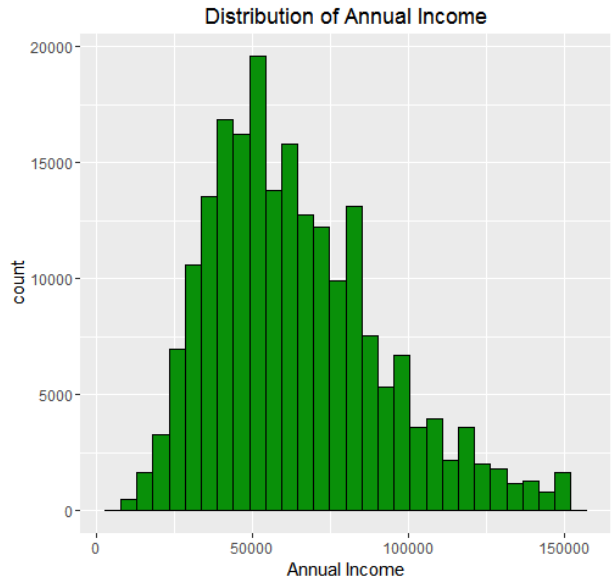
Let’s see the distribution of Loan Amounts Requested, Funded & Investor Funded

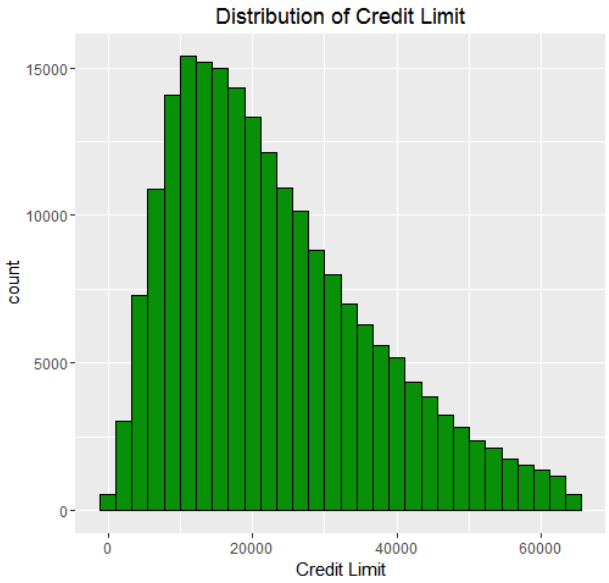






The distributions are quite identical which shows high correlation b/w these three attributes.





1. Annual Income seems to show a normal distribution.

2. Credit limit seems to show an early spike.

Binning Data

Variable X12 is binned and two new variables are created (Interest_rate & Annual Income) which ares binned with the data of X1.

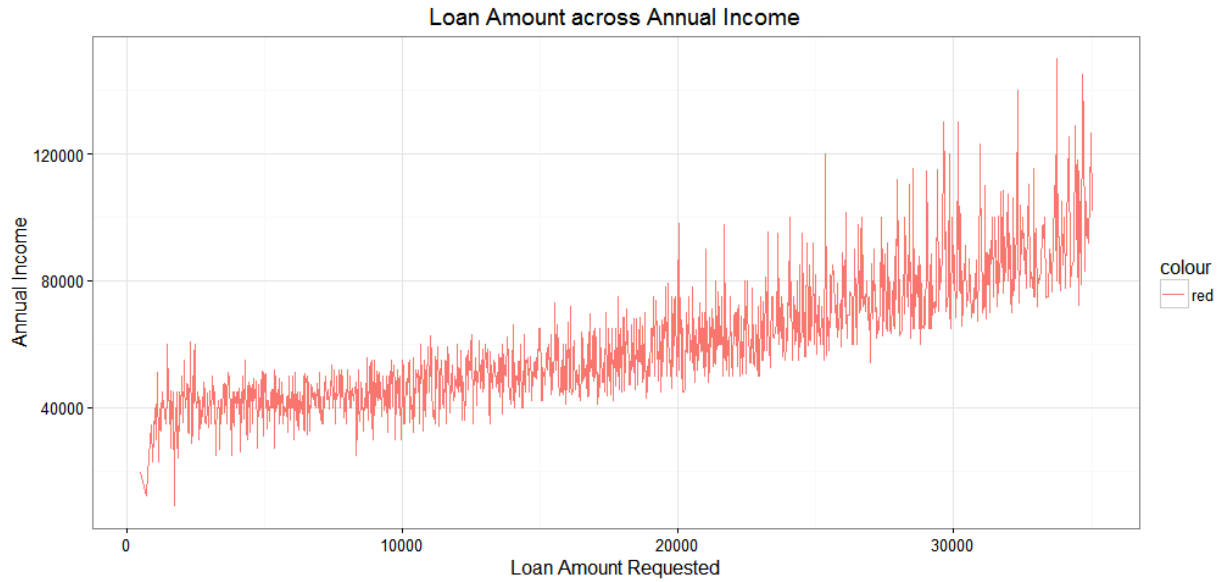
Outliers are removed using box plot.

Multi co linearity Test

```
> vif(data[,c(1,4:6,13,21,22,24,27,28:31,33)])
      Variables        VIF
1            X1   1.402035
2            X4 514.588210
3            X5 635.125113
4            X6 109.039278
5           X13   1.674051
6           X21   1.392660
7           X22   1.062709
8           X24   1.144359
9           X27   1.916908
10          X28   1.057645
11          X29   7.741011
12          X30   3.887233
13          X31   1.865194
14 credit_limit   6.517895
```

```
> vifcor(data[,c(1,4:6,13,21,22,24,27,28:31,33)], th=0.9)
2 variables from the 14 input variables have collinearity problem:

X5 X6

After excluding the collinear variables, the linear correlation coefficients ranges between:
min correlation ( X24 ~ X21 ):  -0.001409212
max correlation ( credit_limit ~ X29 ):  0.7268575

---------- VIFs of the remained variables --------
      Variables      VIF
1            X1 1.362560
2            X4 1.621316
3           X13 1.581624
4           X21 1.385866
5           X22 1.060958
6           X24 1.133631
7           X27 1.938382
8           X28 1.066601
9           X29 7.832404
10          X30 3.807664
11          X31 1.867924
12 credit_limit 6.610032
```

After the muticollinearity test we see that X5 & X6 are having this problem. So they are not fed to the model to avoid any inflation in accuracy of the model.

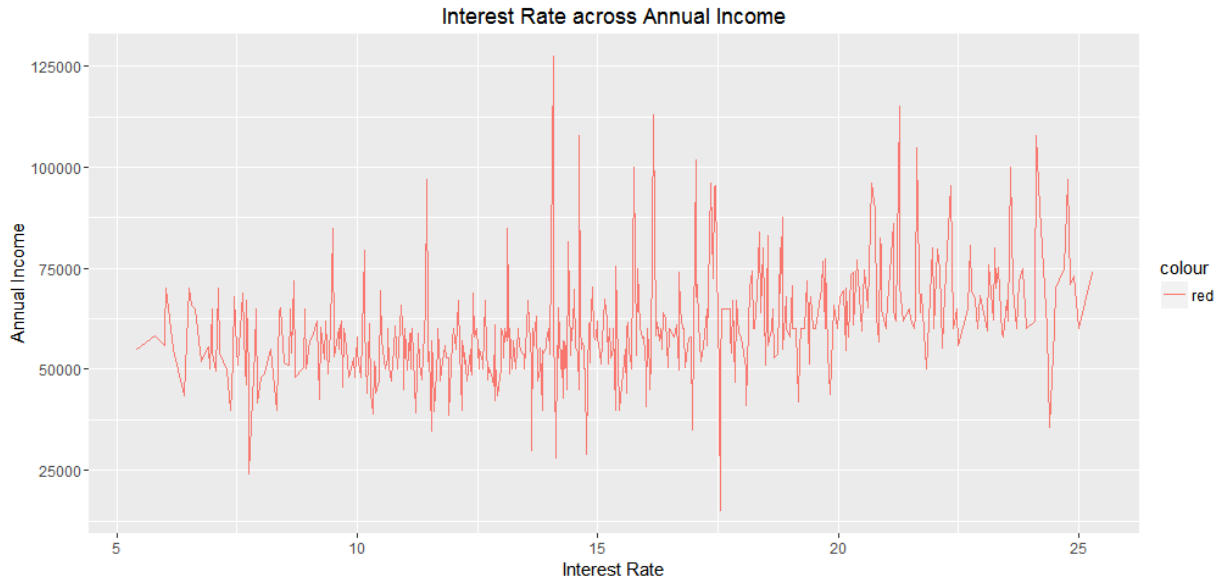


This plot shows that as Annual Income increases Loan Amount also increases but not that marginally.

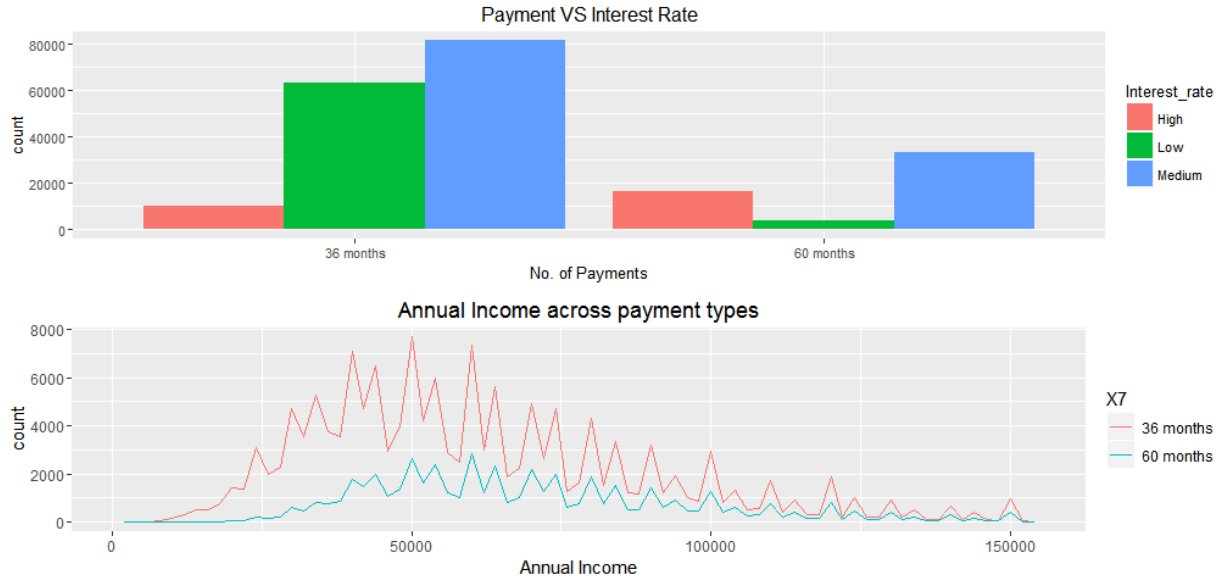
There is relationship between them but not directly linked.

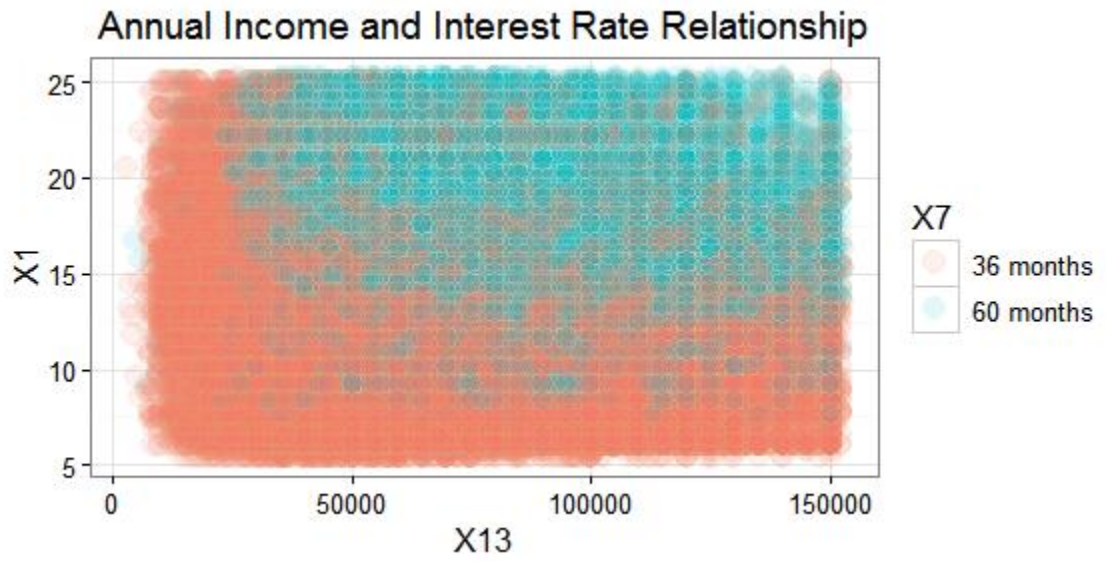
```
> table(data$Annual_income)

 High Income  Low Income Middle Income
       49501        52785       106020

> 49501/208306*100  #Percentage of High Income Group
[1] 23.7636
> 52785/208306*100  #Percentage of Low Income Group
[1] 25.34012
> 106020/208306*100  #Percentage of Middle Income Group
[1] 50.89628
```

From the calculations we see that 50% of the data consists of Middle Income Group. The rest data is divided in half between Low & High Income groups.



This plot as such shows no link between Annual Income & Interest Rate.





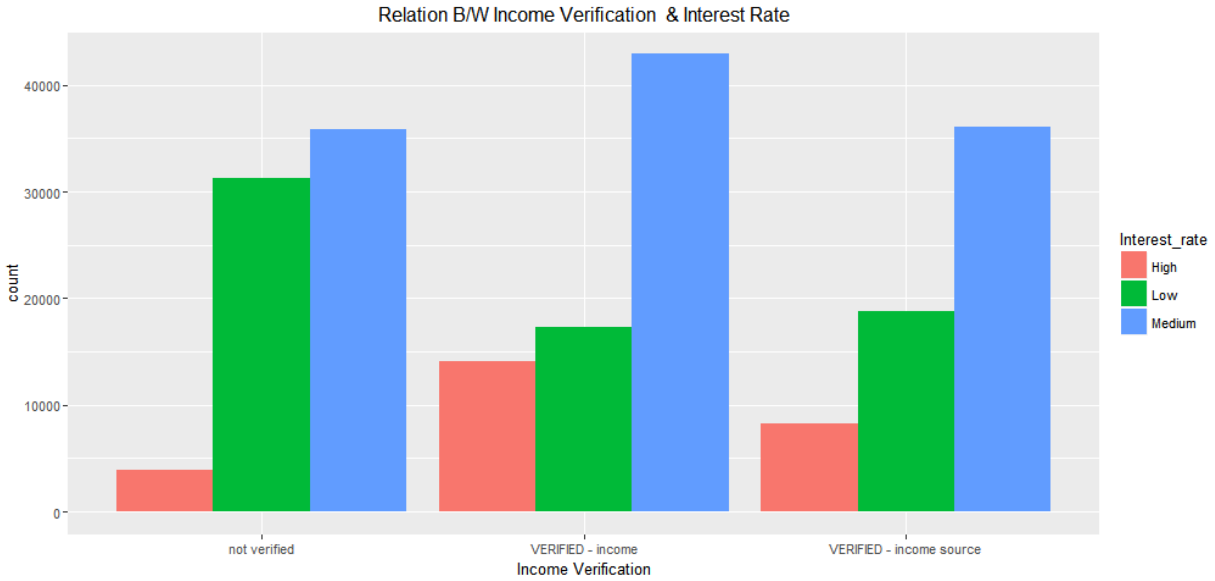
From this plot we see that more people choose for 36 months payment method over 60 months. As 60 months payment method increases the chance of getting high Interest Rate which is shown in second plot.

```
> table(data$X7,data$Interest_rate)

                High   Low    Medium
 36 months   9840   63458  81678
 60 months  16399  3855    33076
```

```
> table(data$X7)

 36 months  60 months
   154976      53330
```

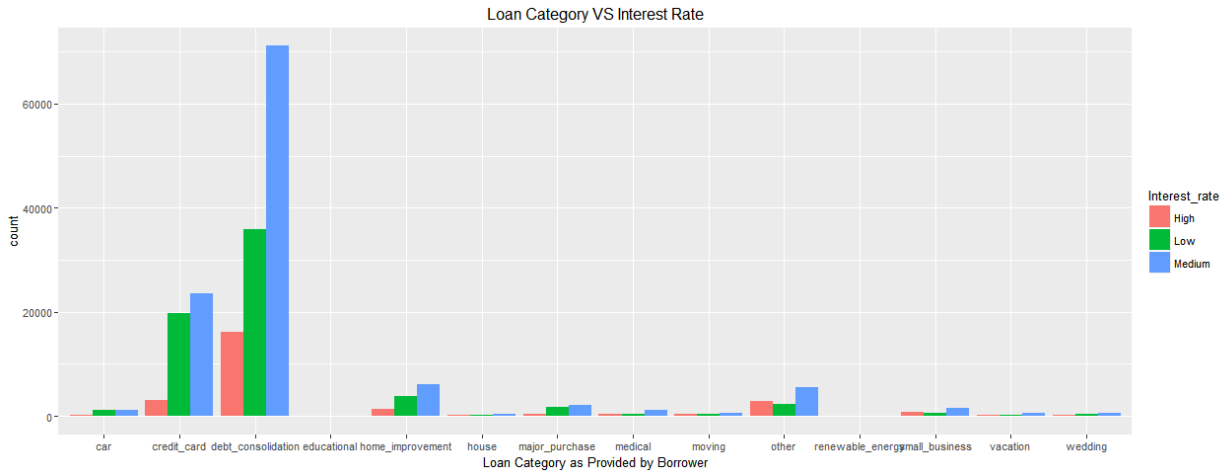
```
> 16399/53330*100
[1] 30.75005

> 9840/154976*100
[1] 6.34937
```

From the calculations we see that there is 30% chance of getting a high Interest Rate for 60 months payment method as compared to 6% for 30 months.

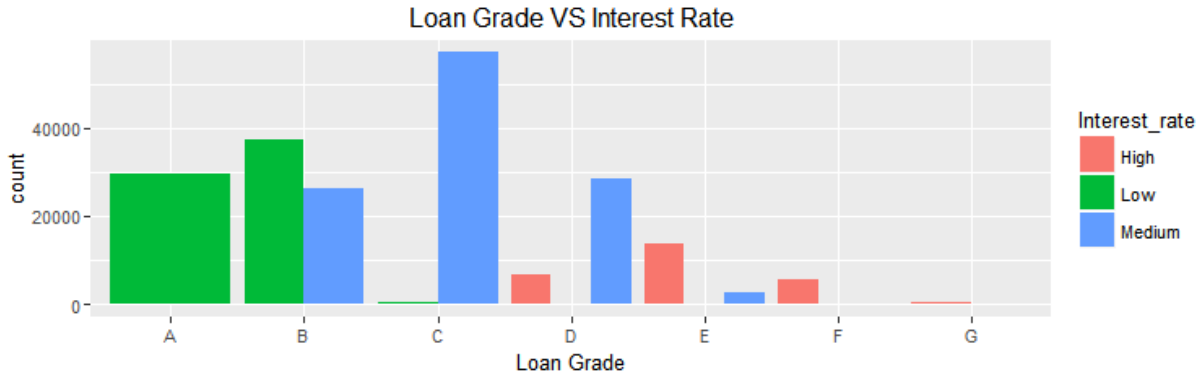


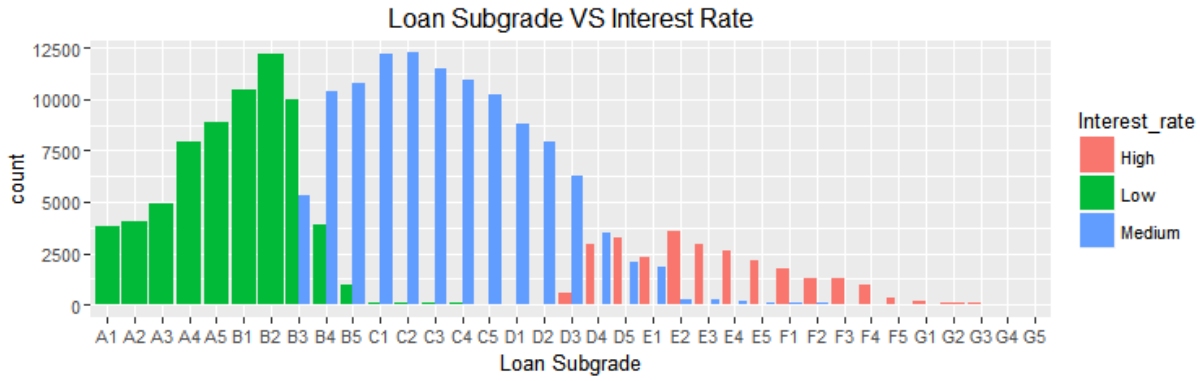
From the plot we see that Income Verification have little impact on Medium Interest where as non verified has the lowest High Interest payers & Highest Low Interest payers among the other two groups. As calculated below:

```
> table(data$X14)

    not verified VERIFIED - income VERIFIED - income source
           70904            74315                    63087


> table(data$X14,data$Interest_rate)

                           High   Low Medium
 not verified              3887 31216  35801
 VERIFIED - income        14131 17281  42903
 VERIFIED - income source  8221 18816  36050

> 3887/70904*100   #Percentage of Not Verified Income having High Interest
[1] 5.48206
> 14131/74315*100  #Percentage of Verified Income having High Interest
[1] 19.015
> 8221/63087*100   #Percentage of Verified Income Source having High Interest
[1] 13.03121


> 31216/70904*100  #Percentage of Not Verified Income having Low Interest
[1] 44.02572
> 17281/74315*100  #Percentage of Verified Income having Low Interest
[1] 23.25372
> 18816/63087*100  #Percentage of Verified Income Source having Low Interest
[1] 29.82548
```



From this plot we see that people generally take loan to pay their credit card debt and to consolidate their debts.

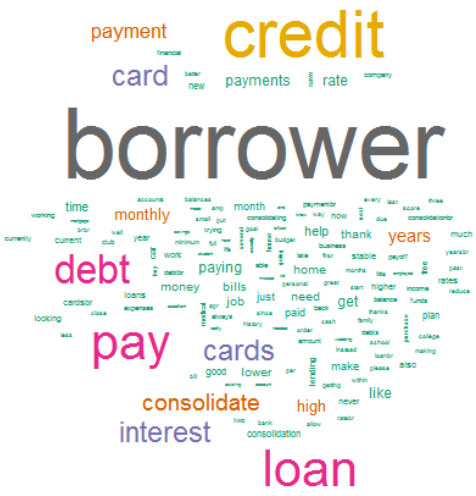




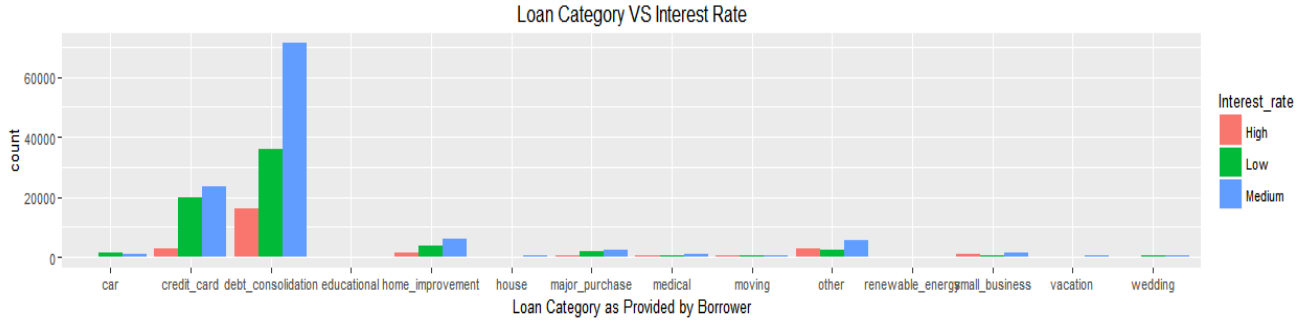
From the above plot we see that as the Loan grade increases Interest rates also increases.

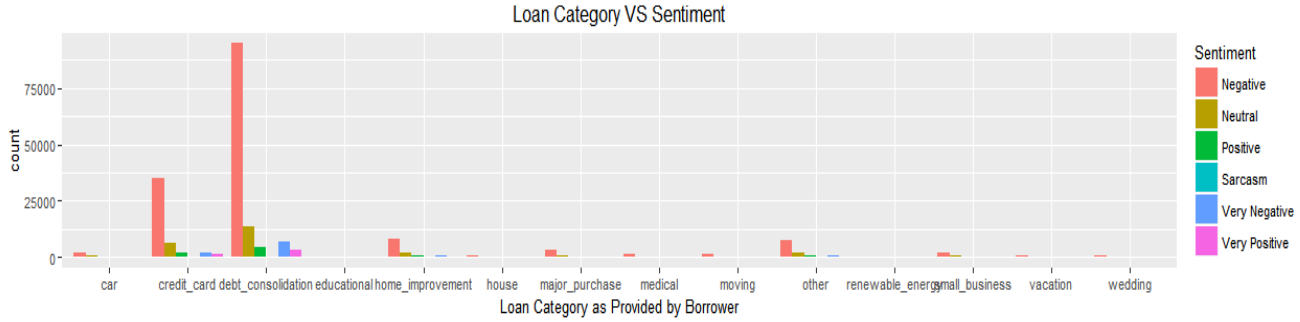
Text Mining

Text Mining done on X16 and word cloud plotted.



Word Cloud on negative sentiments





From this we see that people generally take loan to pay off their debts which explains their negative sentiments.

Model

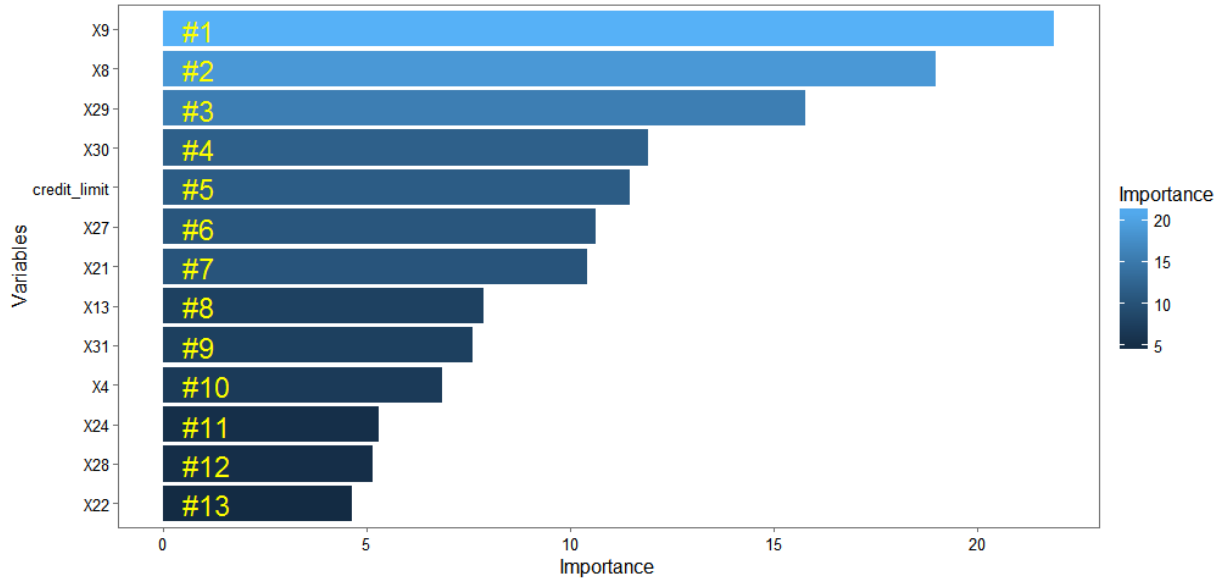
A Random Forest model is used to predict the Interest Rate on test data with 95.8% accuracy.

> rf_model

Call:
randomForest(formula = X1 ~ X4 + X8 + X9 + X13 + X21 + X22 + X24 + X27 + X28 + X29 + X30 + X31 + credit_limit, data = train, importance = TRUE, ntree = 100)
Type of random forest: regression
Number of trees: 100
No. of variables tried at each split: 4

Mean of squared residuals: 0.6894752
% Var explained: 96.09

Variable Importance



From this we see that variable X8, X9 & X29 are good predictors for our response variable X1.

Pros:-

1. For a large number of variables this tree based model is perfect.

2. Gives better model performance as it decreases the variance of the model without increasing the bias.

Cons:-

1. Gets bias for factor variables which shows more levels so variable importance scores are not reliable and we are using two of them in this model.

2. It gives good accuracy but at the expense of high computation time.

Another Model

A Linear Regression model is used and we got an accuracy of 95.7%

Call:
lm(formula = X1 ~ X4 + X9 + X13 + X21 + X22 + X24 + X27 + X28 +
    X29 + X30 + X31 + credit_limit, data = train)

Residuals:

Min	1Q	Median	3Q	Max
-11.3697	-0.5261	0.1144	0.5156	2.3252

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.814e+00	5.700e-02	102.010	< 2e-16	***
X4	3.027e-06	9.640e-07	3.140	0.001692	**
X9A2	5.718e-01	6.001e-02	9.527	< 2e-16	***
X9A3	1.414e+00	5.706e-02	24.783	< 2e-16	***
X9A4	1.788e+00	5.227e-02	34.217	< 2e-16	***
X9A5	2.631e+00	5.154e-02	51.050	< 2e-16	***
X9B1	3.694e+00	5.060e-02	73.004	< 2e-16	***
X9B2	4.693e+00	5.035e-02	93.200	< 2e-16	***
X9B3	5.634e+00	4.903e-02	114.906	< 2e-16	***
X9B4	6.386e+00	4.979e-02	128.250	< 2e-16	***
X9B5	6.992e+00	5.060e-02	138.203	< 2e-16	***
X9C1	7.587e+00	5.081e-02	149.311	< 2e-16	***
X9C2	8.187e+00	5.084e-02	161.046	< 2e-16	***
X9C3	8.781e+00	5.134e-02	171.016	< 2e-16	***
X9C4	9.288e+00	5.166e-02	179.812	< 2e-16	***
X9C5	9.906e+00	5.212e-02	190.070	< 2e-16	***
X9D1	1.056e+01	5.348e-02	197.459	< 2e-16	***
X9D2	1.121e+01	5.408e-02	207.213	< 2e-16	***
X9D3	1.176e+01	5.545e-02	212.045	< 2e-16	***
X9D4	1.231e+01	5.630e-02	218.613	< 2e-16	***
X9D5	1.296e+01	5.787e-02	224.004	< 2e-16	***
X9E1	1.344e+01	6.176e-02	217.567	< 2e-16	***
X9E2	1.400e+01	6.212e-02	225.414	< 2e-16	***
X9E3	1.470e+01	6.518e-02	225.518	< 2e-16	***
X9E4	1.543e+01	6.923e-02	222.846	< 2e-16	***
X9E5	1.610e+01	7.329e-02	219.671	< 2e-16	***
X9F1	1.698e+01	7.939e-02	213.834	< 2e-16	***
X9F2	1.743e+01	8.392e-02	207.751	< 2e-16	***
X9F3	1.763e+01	8.471e-02	208.153	< 2e-16	***
X9F4	1.821e+01	9.782e-02	186.216	< 2e-16	***
X9F5	1.697e+01	1.391e-01	121.935	< 2e-16	***
X9G1	1.748e+01	2.012e-01	86.913	< 2e-16	***
X9G2	1.635e+01	2.891e-01	56.555	< 2e-16	***
X9G3	1.667e+01	5.720e-01	29.148	< 2e-16	***
X9G4	1.720e+01	2.729e-01	63.001	< 2e-16	***
X9G5	1.831e+01	5.721e-01	32.004	< 2e-16	***
X13	-7.003e-07	2.609e-07	-2.684	0.007271	**
X21	3.308e-05	8.839e-06	3.742	0.000183	***

```
X22           1.009e-02  7.314e-03   1.379 0.167784
X24           2.212e-02  5.812e-03   3.807 0.000141 ***
X27           7.712e-03  1.810e-03   4.262 2.04e-05 ***
X28          -3.936e-02  1.245e-02  -3.161 0.001572 **
X29           1.610e-06  1.888e-06   0.853 0.393895
X30           2.478e-03  4.836e-04   5.124 3.02e-07 ***
X31          -1.224e-03  7.748e-04  -1.580 0.114228
credit_limit  1.162e-08  1.079e-06   0.011 0.991410
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.806 on 19954 degrees of freedom
Multiple R-squared:  0.9633,    Adjusted R-squared:  0.9632
F-statistic: 1.162e+04 on 45 and 19954 DF,  p-value: < 2.2e-16
```

From this we see that Variable X9 is a very important predictor along with X1, X24, X27 & X30. Variance explained by this model 96.32% which shows the model quality.

Pros:-

1. Computation time is quiet low as compared to Random Forest.
2. We can test how the model fits using R square & Adjusted R square.
3. Gives statistical significance for variable importance.

Cons:-

1. A lot of pre-processing required before it fed to a model.
2. Accuracy is affected if variables of right orders are not added.

Conclusion:-

A random forest will be a better choice as it is tree based method if we can bear the computation time.