Final Project Report

Prediction of Lending Club Interest Rates

OR /SYST 438/538: Fall 2016

Professor: K C Chang

Team-06

Venkata Jayandra Kumar Lade — Sasidhar Thapiti

G01046700 — G00989375

Table of Contents

Problem Statement:

Lending Club is the world's largest online credit marketplace, connecting borrowers and investors, which gives loans for different purposes. The interest rate may vary with certain attributes like Annual Income, Loan Amount, Home Ownership Status, Purpose of loan, Previous History of Debits, etc. The purpose of this project is to predict the interest rates of the lending club and to find out which factors plays major role.

Project Goal:

To predict the interest rates of Lending club using regression analysis method.

Data Collection:

Data has been collected from the official website of Lending club and the link has been provided below.

https://resources.lendingclub.com/LoanStats3c.csv.zip#sthash.TKf5NGkh.dpuf

It has the information of several attributes which affect to the interest rates like Annual Income, Loan Amount, Home Ownership Status, Purpose of loan, Previous History of Debits, etc. The attributes are of two types in this data set i.e., categorical and numerical.

The response variable is int_rate and remaining all are predictors.

Data Pre-processing:

The data which has obtained requires some pre-processing to perform analysis. The things which are done in pre-processing is:

- Eliminating unnecessary attributes, like id, address etc.
- Eliminating rows which has null values
- Formatting the Data, like
    - Removing months in term column
    - Changing percentages into numerical values (ex: 12% to 0.12)

Technical Approach:

Preliminary Analysis:

Before the model selection process, the dataset has been explored using visual tools. In doing so, we could get a better understanding of how the data are distributed and how the predictor variables are related to the response variable, is calculated using correlation. Each variable in the dataset was plotted using histograms. From this, we could determine which variables might require transformation. All this analysis can be conducted on numeric variables.

From these histograms (see appendix) we can determine that many variables are skewed. This tells us that transformation might be necessary. And There are some variables which are relatively normal, and may not require transformation. We continued our preliminary analysis by plotting using correlation of each predictor variable against the response and other predictors. Below we are showing the correlation plot against each variable.

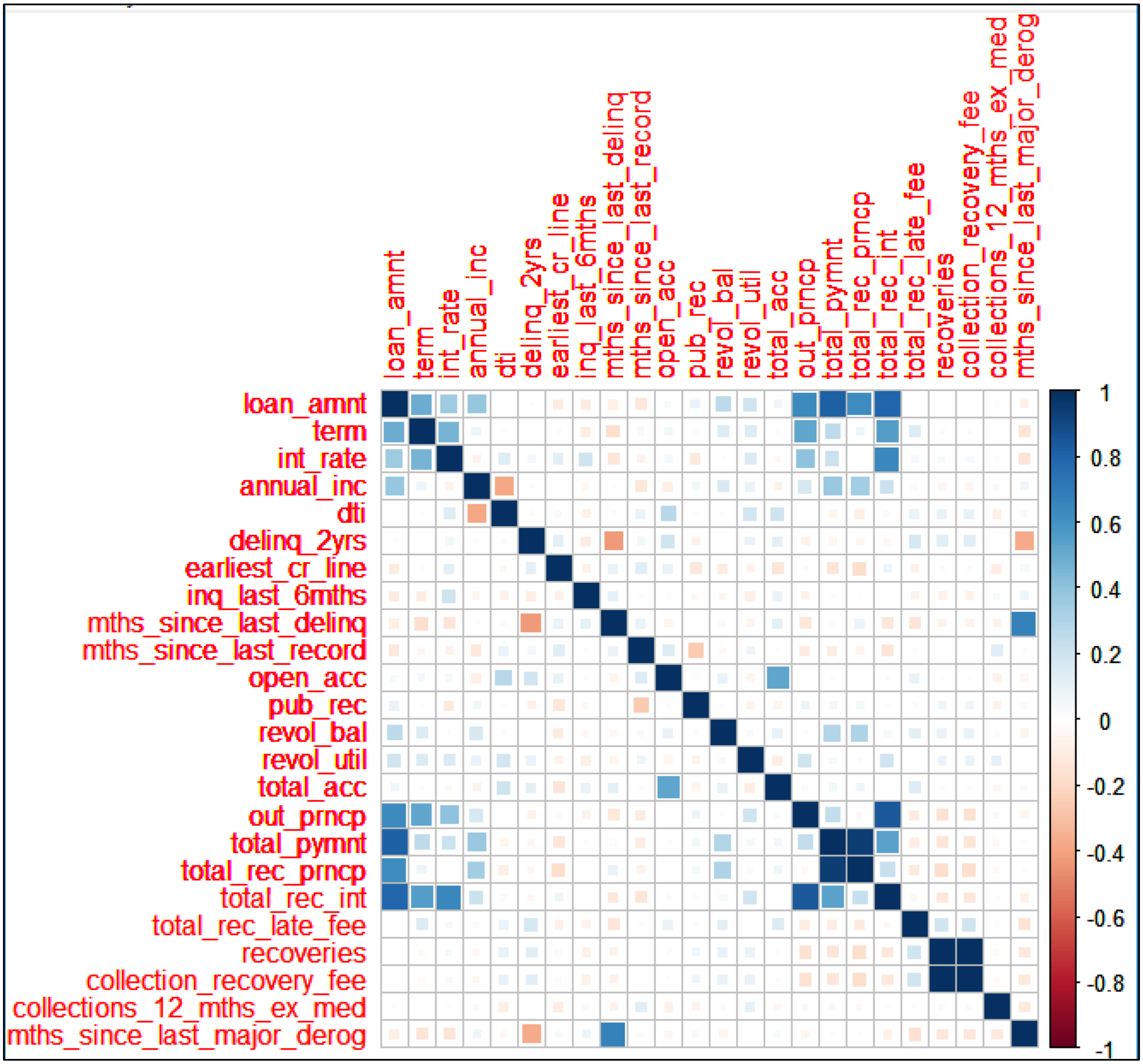


Figure 1: Correlation plot among the variables including predictors and response variable

From the above correlation plot, we can determine that many of the predictors are strongly related with the response variable. We can see that there is high potential for a multicollinearity issue in this data. A large portion of the variables are highly correlated with one another. We should be especially concerned with correlations larger than $r = \pm 0.80$ amongst two predictor variables.

Now that we have an idea about how the data is distributed and about the relationships amongst the variables, we can proceed with the model selection process.

Multiple Linear Regression:

Multiple linear regression is a method, which is performed to predict a variable. It is used to predict a variable which is called as response variable with the help of two or more variables which are called predictors. The general equation is shown in below:

$$Y = \alpha + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \ldots + \beta iXi$$

Where,

- ➔ Y = Predicted Variable or Dependent Variable
- ➔ α = Intercept of the Regression Equation
- ➔ X1, X2….Xi = Predictor Variables or Independent Variables
- ➔ β1, β2, β3…βi = Regression Coefficients

Here, we have conducted step wise regression which are backward, forward and both.

Backward Elimination:

Backward elimination, which involves starting with all candidate variables, testing at each step and the deletion of each variable which eventually gives a best model.

Forward Elimination:

Forward elimination, which involves starting with no candidate variables, testing at each step and the addition of each variable which eventually gives a best model.

Both Elimination:

Both elimination, is combined of both forward and backward elimination, testing at each step for variables to be included or excluded.

The summary of all models is shown in appendix.

Model Selection:

AIC & BIC:

The best model is selected using AIC (Akaike information criterion) and BIC (Bayesian information criterion) ,which have smaller AIC and BIC values.

AIC

	df	AIC
backward	21	-535.5156
forward	11	-536.8188
both	21	-535.5156

BIC

	df	BIC
backward	21	-476.9783
forward	11	-506.1564
both	21	-476.9783

Interpreting above values, we can conclude that forward model is best.

Variance Inflation Factors:

By looking at VIF values, we can say that whether any transformation is required or not.

```
> vif(forward)
    total_rec_int        annual_inc         out_prncp         loan_amnt              term 
         6.990780          1.240432          4.414256          3.382856          1.739687 
   inq_last_6mths total_rec_late_fee           pub_rec        revol_util 
         1.050961          1.082770          1.028658          1.074150 
```

However, looking at the VIF values, we can see that there is no multicollinearity issue, as all corresponding VIF values are smaller than 10.00. So, there is no need of any transformations.

From the diagnostic plots for this model (see appendix), it appears that all the MLR assumptions are supported. We can run formal tests to check these assumptions further.

Formal Linear Assumptions Tests:

Linearity:

To check the linearity, we have plotted residuals vs fitted. The plot is shown below:

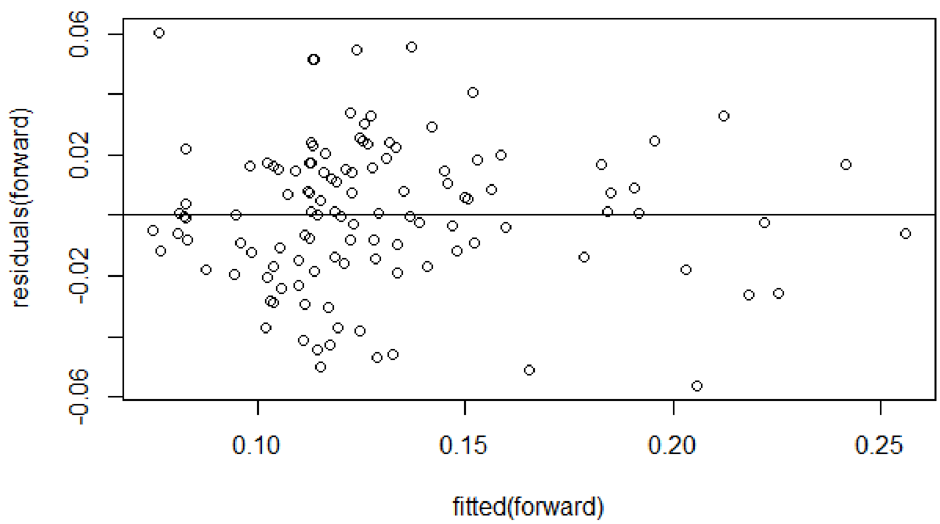


Figure 2: Residuals Vs fitted for the forward model

From the above graph, we can say that there is an even distribution of points around the centered line. By which we can say that linearity assumption is supported.

Anderson-Darling Test:

Anderson-darling test is a method, by which we can say whether the distribution is normal or not. The formal hypothesis of this test is below:

H_0 (Null Hypothesis): The distribution in normal

H_a (Alternative Hypothesis) : The distribution is not normal

```
> ad.test(forward$residuals)

        Anderson-Darling normality test

data:  forward$residuals
A = 0.3473, p-value = 0.474
```

From the above results, we can say that the distribution is normal because p-values is greater than 0.05 which fails to reject the null hypothesis.

Analysis of Variance(ANOVA):

The anova table is generated using a function called anova in R directly for the model.

```
> anova(forward)
Analysis of Variance Table

Response: int_rate
                    Df   Sum Sq  Mean Sq  F value    Pr(>F)    
total_rec_int        1 0.098397 0.098397 162.2188 < 2.2e-16 ***
annual_inc           1 0.011929 0.011929  19.6658 2.195e-05 ***
out_prncp            1 0.013159 0.013159  21.6936 9.007e-06 ***
loan_amnt            1 0.010298 0.010298  16.9766 7.357e-05 ***
term                 1 0.007842 0.007842  12.9282 0.0004861 ***
inq_last_6mths       1 0.004459 0.004459   7.3512 0.0077781 ** 
total_rec_late_fee   1 0.002816 0.002816   4.6431 0.0333590 *  
pub_rec              1 0.002832 0.002832   4.6691 0.0328789 *  
revol_util           1 0.001417 0.001417   2.3359 0.1292916    
Residuals          110 0.066723 0.000607                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Analysis of Variance

By seeing that table, we can say the factors are significant or not. And we can also calculate the Residuals sum of squares, Regression sum of squares and total sum of squares and R-squared also.

Residuals Sum of Squares:

Residual sum of squares is the sum of squares of deviations of the predicted values from the actual values of dependent variable.

$$\text{RSS} = \sum_{i=1}^{n}\left(\hat{Y} - Y\right)^2$$

Regression Sum of squares:

Regression sum of squares is the sum of squares of deviations of the predicted values from the mean values of dependent variable.

$$\text{RSS} = \sum_{i=1}^{n}\left(\hat{Y} - \bar{Y}\right)^2$$

Total Sum of Squares:

Total sum of squares is the sum of squares of deviations of the actual value from the mean value.

$$\text{TSS} = \sum_{i=1}^{n}\left(Y - \bar{Y}\right)^2$$

R-squared value:

R-squared= Regression S.S/Total S.S = 1- Residual S.S/Total S.S

- **Regression Sum of Squares:** 0.1537
- **Residual Sum of Squares:** 0.06672
- **Total Sum of Squares:** 0.2205
- **R-squared : 0.697**

Multiple Linear Regression Equation:

The equation of our model is shown below:

Int_rate = 0.076 + 2.75 * 10^-5 total_rec_int – 1.13 * 10^-7 annual_inc – 4.95 * 10^-6 out_prncp – 2.315 * 10^-6 loan_amnt + 1.209 * 10^-3 term + 5.802 * 10^-3 inq_last_6mths – 1.572 * 10^-3 total_rec_late_fee – 5.866 * 10^-3 pub_rec + 1.645 * 10^-2 revol_util

Int_rate: Interest rate on the loan

total_rec_int: Interest received to the date

annual_inc: The self reported annual income

out_prncp: Remaining outstanding principal

loan_amnt: The listed amount of the loan applied by the borrower

term: Number of payments on the loan

inq_last_6mths: Number of inquiries in past 6 months

total_rec_late_fee: Total late fees received to the date

pub_rec: Number of derogatory public records

revol_util: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
The predictors in above equations plays major role in predicting.

Summary and Conclusion:

To know how well our model is working, we have predicted specifically for two rows from the validation set which are 1st and 59th row, which are predicted as 0.119 and 0.126 respectively while the original values are 0.104 and 0.114 respectively. Based on this we can say that we can rely on this model(forward). Here, we have calculated error using root mean square values, which preforms square root of mean of square of difference between all predicted values and originals values in validation set. For which we got a value of 0.03.

In conclusion, we can say that we can predict the interest rates of lending club with error of ±0.03 to the original values.

References:

[1] Analytics for Financial Engineering and Econometrics lectures by KC Chang

[2] Statistics and Data Analysis for Financial Engineering by David Ruppert

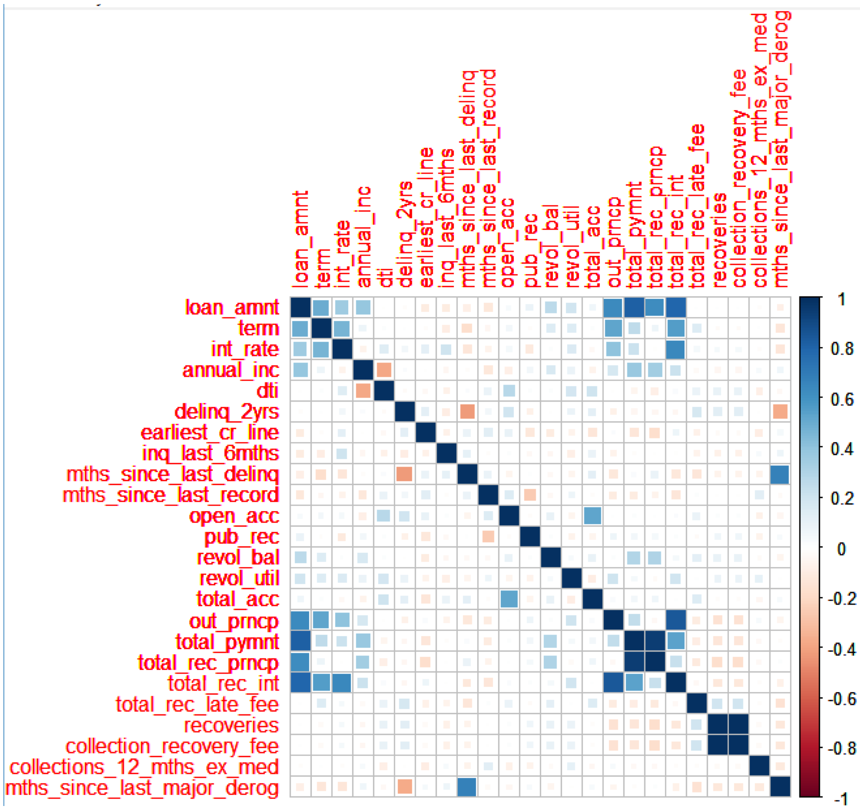
[3] https://resources.lendingclub.com/LoanStats3c.csv.zip#sthash.TKf5NGkh.dpuf

[4] https://en.wikipedia.org/wiki/Stepwise_regression

[5] http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/model-assumptions/what-is-a-variance-inflation-factor-vif/

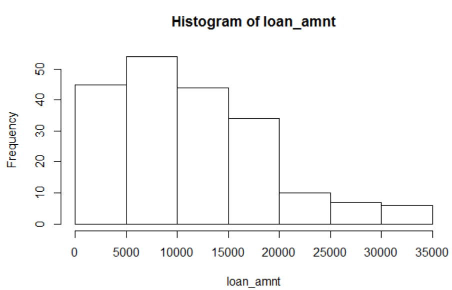
Appendix:

```
#reading data set

d1=read.csv("LoanStats3c_updated.csv",header=TRUE)

#converting all the null spaces into NA

d1[d1==""]<-NA

d1<-na.omit(d1)

#removing months in term

s1 <- sapply(strsplit(as.character(d1[,"term"])," "), function(x) (x[2]))

s1 <- as.integer(s1)

s1<-data.frame(s1)

d1[,"term"]<-s1

d1=as.data.frame(d1)

#preliminary analysis

#selecting numerical variables from the data set

for(i in 1:ncol(d1))

{

  if(i==1)

    tf<-is.numeric(d1[,i])

  else

    tf[i]<-is.numeric(d1[,i])

}

colnum_data<- which(tf)

d4=d1[,colnum_data]

m=cor(d4)

library(corrplot)

corrplot(m,method = "square")
```

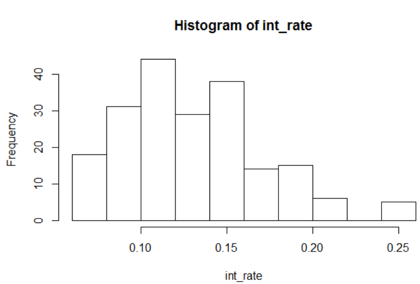


#individual Histograms

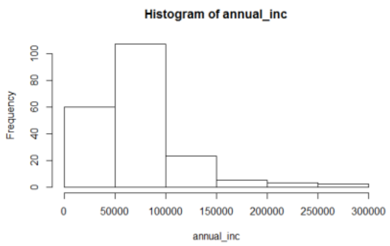
hist(loan_amnt) *#skewed*



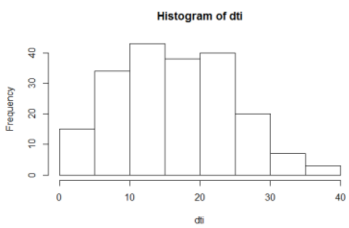
hist(int_rate) *#skewed*



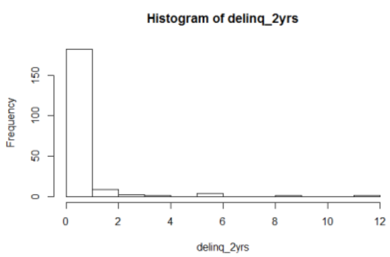
hist(annual_inc) *#skewed*



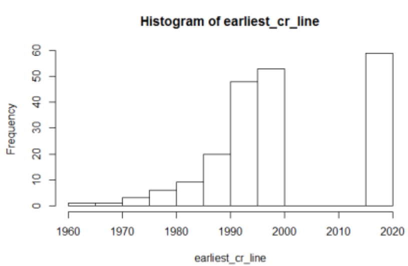
hist(dti)



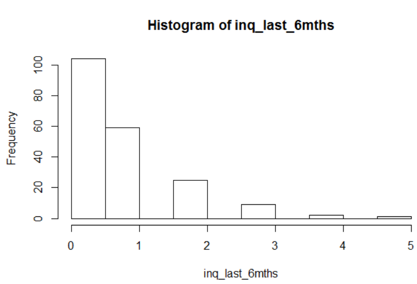
hist(delinq_2yrs) *#skewed*



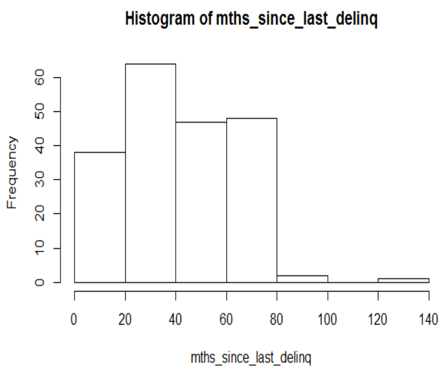
hist(earliest_cr_line) *#skewed*



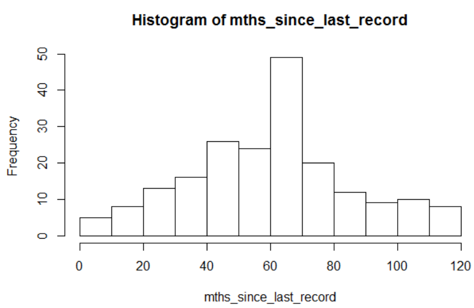
hist(inq_last_6mths) *#skewed*



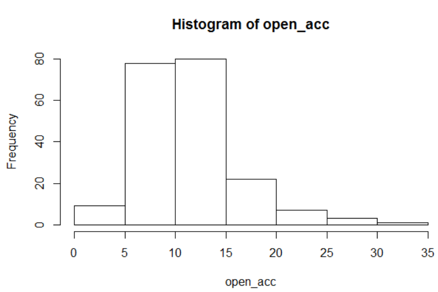
hist(mths_since_last_delinq) *#skewed*



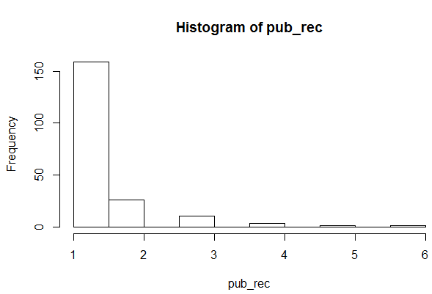
hist(mths_since_last_record)



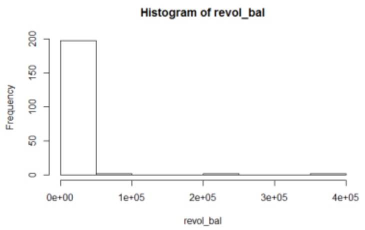
hist(open_acc) *#skewed*



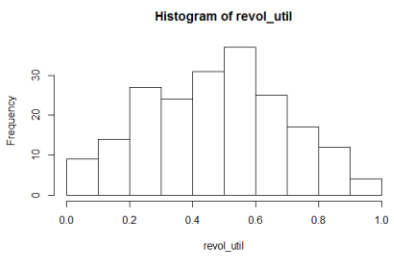
hist(pub_rec) *#skewed*



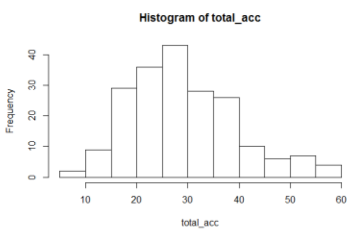
hist(revol_bal) #*skewed*



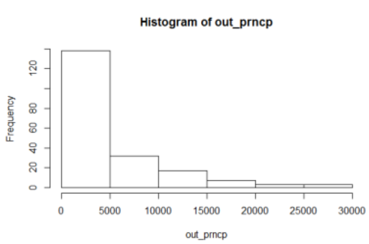
hist(revol_util)



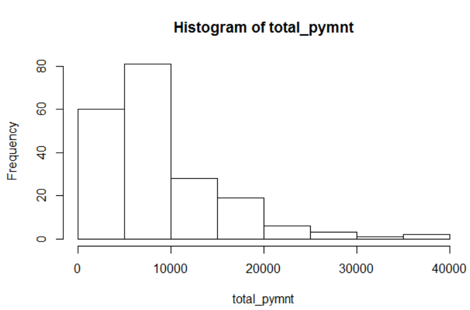
hist(total_acc) #*skewed*



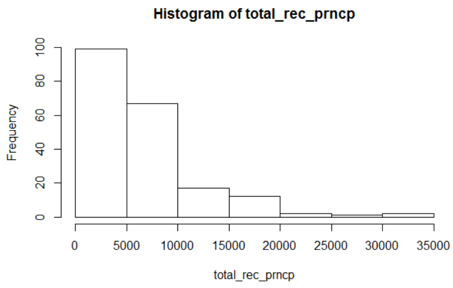
hist(out_prncp) #*skewed*



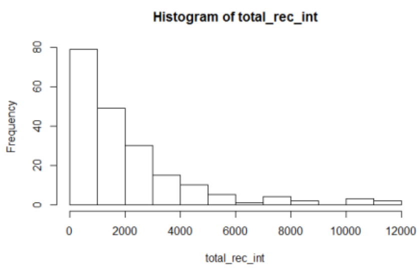
hist(total_pymnt) #*skewed*



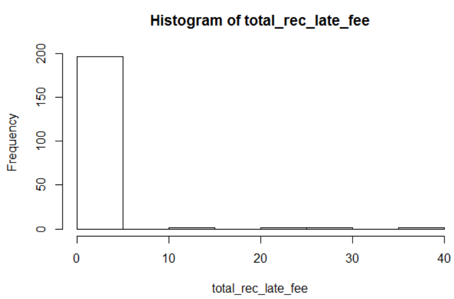
hist(total_rec_prncp) *#skewed*



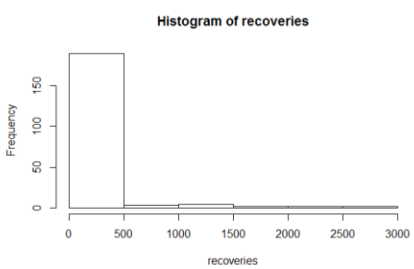
hist(total_rec_int) *#skewed*



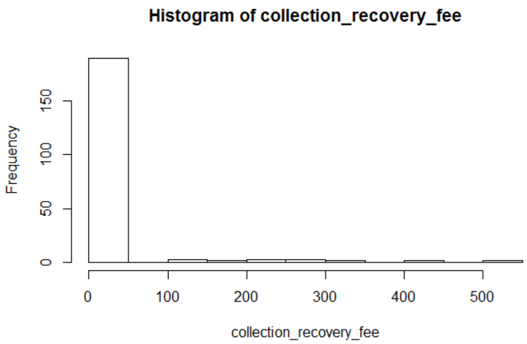
hist(total_rec_late_fee) *#skewed*



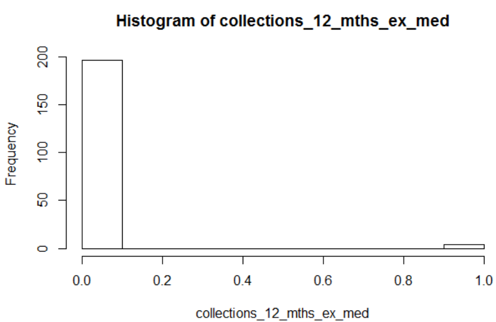
hist(recoveries) *#skewed*



hist(collection_recovery_fee) *#skewed*

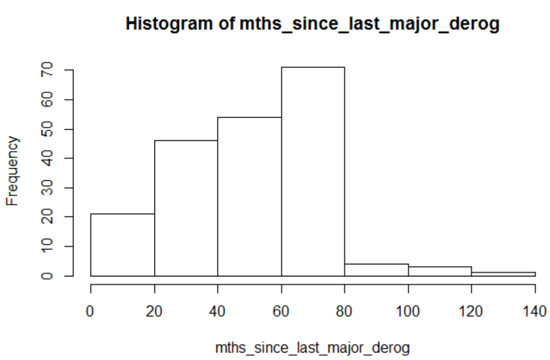


hist(collections_12_mths_ex_med) *#skewed*



---

hist(mths_since_last_major_derog) *#skewed*



#splitting the data set into training and validation

set.seed(2016)

traindata<-sample(n,n*0.6,replace = FALSE)

training<-d1traindata,]

validation<-d1[-traindata,]

#fitting a model for the data through regression

#backward regression

trainfit_temp<-lm(int_rate~.,data=training)

backward<-step(trainfit_temp,direction = 'backward')

#forward regression

null<-lm(int_rate~1,data=training)

forward<-step(null, scope=list(upper=trainfit_temp), direction='forward')

#both direction regression

both<-step(trainfit_temp, direction='both')

#AIC and BIC values for all model—model selection

AIC(backward,forward,both)

BIC(backward,forward,both)

```
> AIC(backward,forward,both)
         df       AIC
backward 21 -535.5156
forward  11 -536.8188
both     21 -535.5156
> BIC(backward,forward,both)
         df       BIC
backward 21 -476.9783
forward  11 -506.1564
both     21 -476.9783
```

#Performing VIF to check multi collinearity

library(HH)

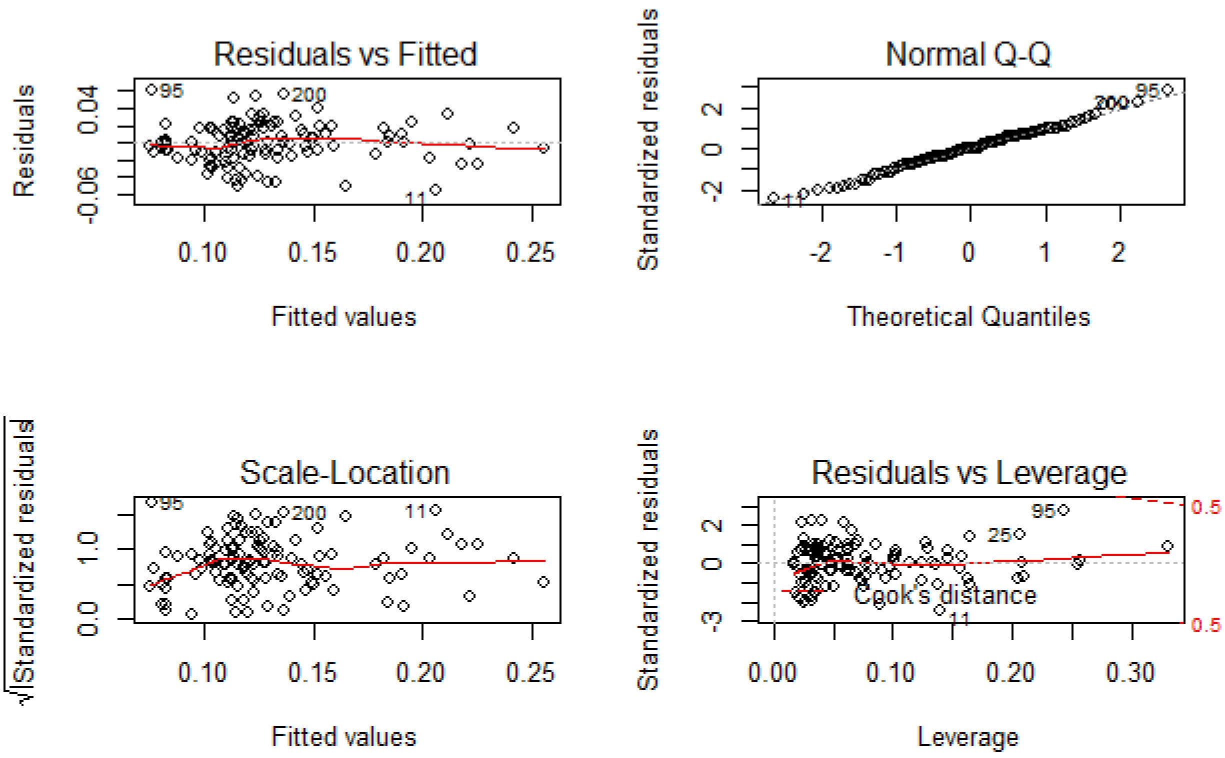
vif(forward)

```
> vif(forward)
    total_rec_int        annual_inc         out_prncp         loan_amnt              term
         6.990780          1.240432          4.414256          3.382856          1.739687
    inq_last_6mths total_rec_late_fee           pub_rec        revol_util
         1.050961          1.082770          1.028658          1.074150
```

#Diagnostic plot

par(mfrow=c(2,2))

plot(forward)

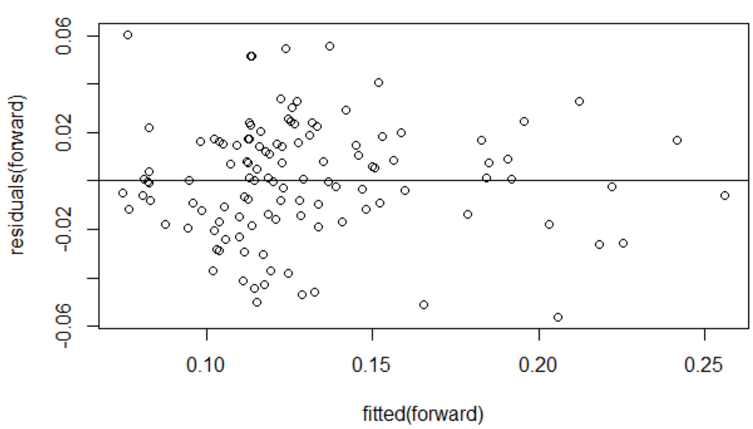


#linearity

dev.off()

plot(fitted(forward),residuals(forward))

abline(0,0)



#Anderson-Darling test

library(nortest)

ad.test(forward$residuals)

```
> ad.test(forward$residuals)

        Anderson-Darling normality test

data:  forward$residuals
A = 0.3473, p-value = 0.474
```

#Analysis of variance

a=anova(forward)

a

```
> anova(forward)
Analysis of Variance Table

Response: int_rate
                  Df   Sum Sq  Mean Sq  F value    Pr(>F)    
total_rec_int      1 0.098397 0.098397 162.2188 < 2.2e-16 ***
annual_inc         1 0.011929 0.011929  19.6658 2.195e-05 ***
out_prncp          1 0.013159 0.013159  21.6936 9.007e-06 ***
loan_amnt          1 0.010298 0.010298  16.9766 7.357e-05 ***
term               1 0.007842 0.007842  12.9282 0.0004861 ***
inq_last_6mths     1 0.004459 0.004459   7.3512 0.0077781 ** 
total_rec_late_fee 1 0.002816 0.002816   4.6431 0.0333590 *  
pub_rec            1 0.002832 0.002832   4.6691 0.0328789 *  
revol_util         1 0.001417 0.001417   2.3359 0.1292916    
Residuals        110 0.066723 0.000607                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#regression sum of squares

sum(a$`Mean Sq`)

```
[1] 0.1537551
```

#residual sum of squares

a$`Sum Sq`[10]

```
[1] 0.06672281
```

#total sum of squares

(sum(a$`Mean Sq`)+a$`Sum Sq`[10])

```
[1] 0.2204779
```

#r-squared

sum(a$`Mean Sq`)/(sum(a$`Mean Sq`)+a$`Sum Sq`[10])

```
[1] 0.6973719
```

#predicting int_rates for entire vildation set

pred=predict(forward,validation)

#root mean square for error caluclation

rms=sqrt(mean((pred-validation$int_rate)^2))

rms

```
[1] 0.0303767
```

#predicting for specific rows

pred[1]

```
0.1198857
```

validation[1,]$int_rate

```
0.1049
```

pred[59]

```
0.1260201
```

validation[59,]$int_rate

```
0.1144
```