MAJOR PROJECT ON

“INTEREST RATE PREDICTION”

BY

ASHISH KUMAR MISHRA

Under the esteemed guidance of Shri. Muquayyar Ahmed,

INDEX

Major Project: Interest Rate Prediction

Objective

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.

Parameters

Variable	Definition
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

INTRODUCTION

X1
This variable defines interest rate on loan amounts.

X2
It consists of unique id for loan.

X3
It consists of unique id for borrower.

X4
It consists of loan amount requested by borrower.

X5
It consists of loan amount funded to borrower.

X6
It consists of investor funded portion of loan amount.

X7
It consists of number of payments done. This is categorical variable. It consists of only two values: 36 and 60.

X8
It consists of different grades of loans.

X9
It consists of different subgrades of loans.

X10
It consists of job title of employer.

X11

It tells about no. of years employed.

X12

It tells about home ownership status as: RENT, OWN, MORTGAGE, OTHER.

X13

It consists of annual income of borrower.

X14

It consists of income and its source status whether verified or not.

X15

It consists of Date when loan vas issued.

X16

It consists of reason for which loan was provided.

X17

It consists of loan category provided by borrower.

X18

It consists of loan title provided by borrower.

X19

It consists of first three no. of zip code.

X20

It consists of state of borrower.

X21

It includes 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

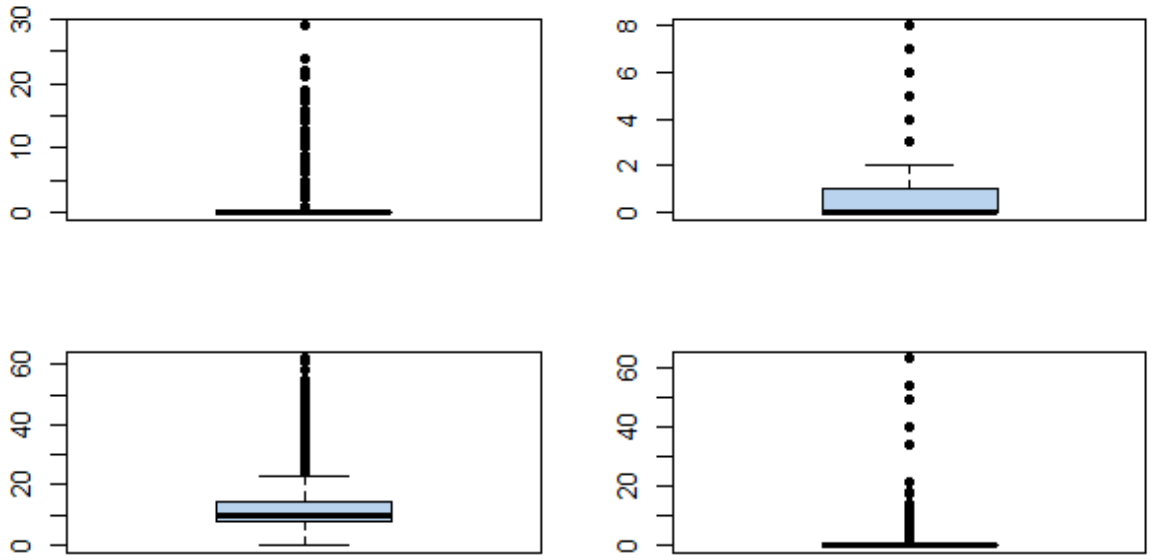
Data Pre-Processing

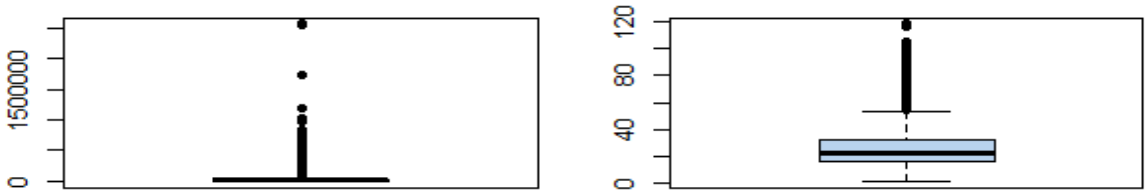
Data is read from csv files and data of both, metadata and Data for Cleaning & Modeling, a data frame is created with Data for Cleaning & Modeling. Data type is changed to numeric for fields imported as factors. All fields except "X1" and character variables are normalized and extreme outliers removed.

Feature Selection

We have got 32 variables in dataset in which first variable is about interest rate and next two are borrower id and loan id's respectively. We have to predict interest rates for a given dataset named "Holdout for Testing".

We have many variables with most of na's and outliers, as shown in picture below:





So above figures shows that our variables have so much outliers. We have to remove it and make it free from outliers.

Major reasons for loan

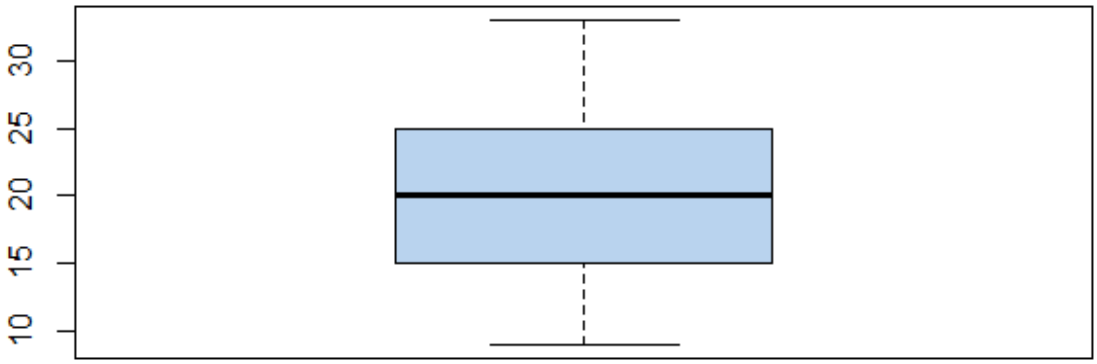
By wordcloud analysis as given below:

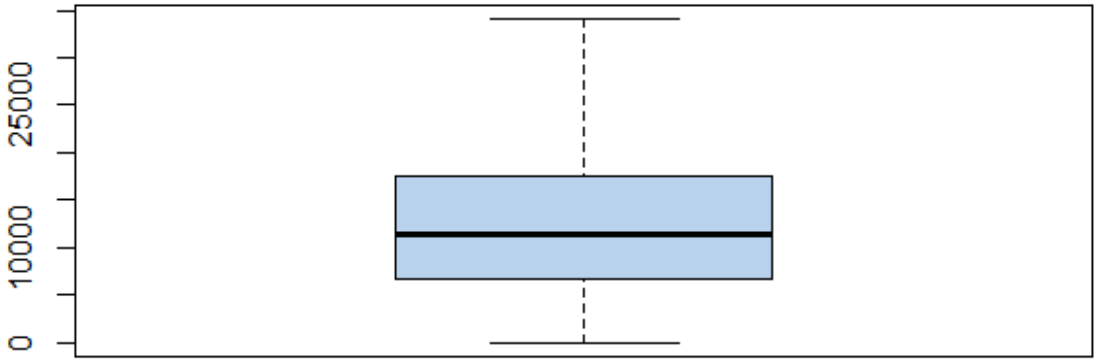
So we have to do some more strings removal and preprocessing on text. After removing those here we found another wordcloud:

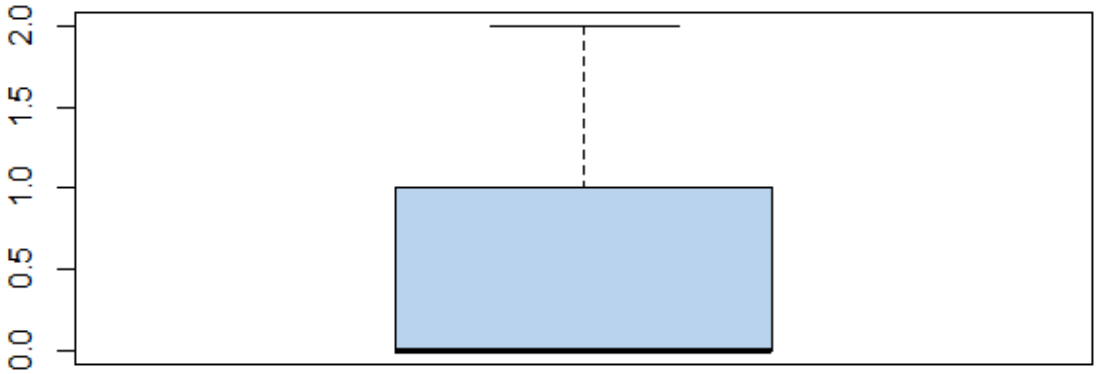
Which shows us that major reasons for debt is credit, loan, interest, consolidate as they are shown big.

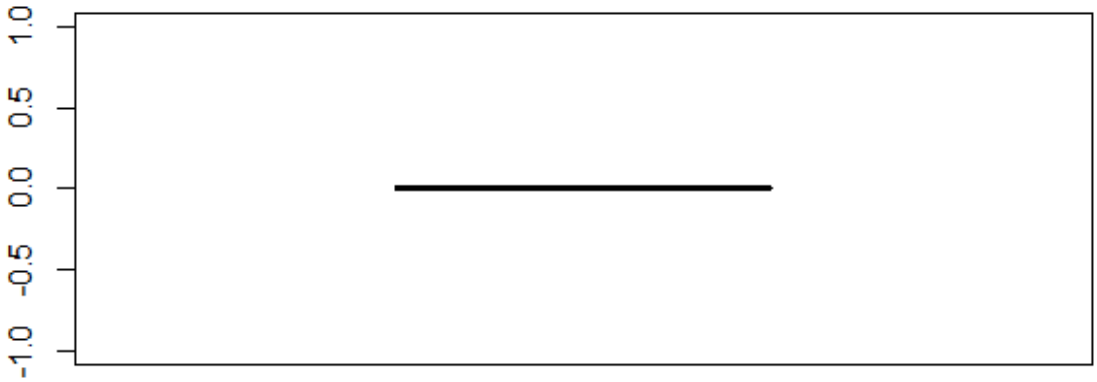
Feature Elimination

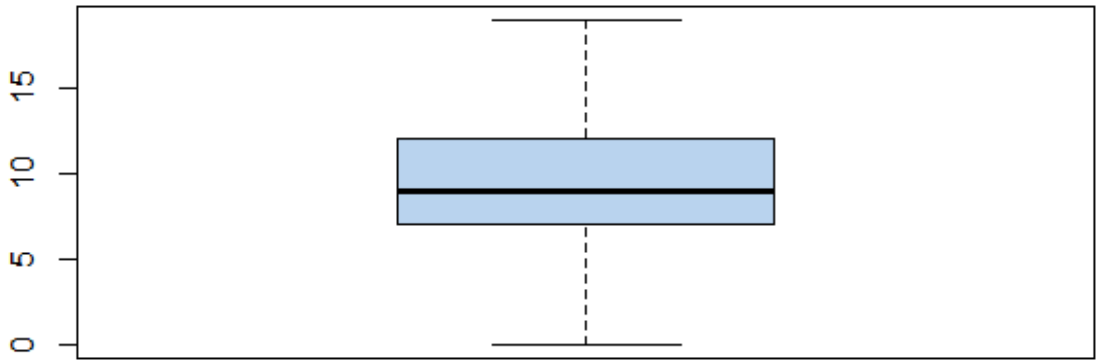
After removing outliers we have found our variables error free from outliers, as given below:


10
15
20
25
30


0
10000
25000


0.0
0.5
1.0
1.5
2.0


-1.0
-0.5
0.0
0.5
1.0



Modeling

As this dataset is very big in size so normal algorithm application is not possible. That is why I have choosed h2o to apply various algorithms on it. I have applied deep learning algorithm on data after dividing in training and test dataset.

Here is some statistics by applying deep learning algorithm:

```
> h2o.performance(dlearning.model)
H2OMultinomialMetrics: deeplearning
** Reported on training data. **
Description: Metrics reported on temporary training frame with 9950 samples

Training Set Metrics:
=====================
Metrics reported on temporary training frame with 9950 samples

MSE: (Extract with `h2o.mse`) 0.03645211
R^2: (Extract with `h2o.r2`) 0.9999978
Logloss: (Extract with `h2o.logloss`) 0.1713691
Mean Per-Class Error: 0.2474252
Confusion Matrix: Extract with `h2o.confusionMatrix(<model>,train = TRUE)`)
=========================================================================
Confusion Matrix: vertical: actual; across: predicted
         V1 10.00% 10.01% 10.08% 10.14% 10.15% 10.16% 10.20% 10.25% 10.28% 10.33
       1459      0      0      0      0      0      0      0      0      0
10.00%    1      6      0      0      0      0      0      0      0      0
10.01%    0      0      0      0      0      0      0      0      0      0
10.08%    0      0      0      0      0      0      0      0      0      0
10.14%    0      0      0      0      0      0      0      0      0      0
```

It took almost 5hrs to run this algorithm and We see improvement in the R^2 metric as in above figure.

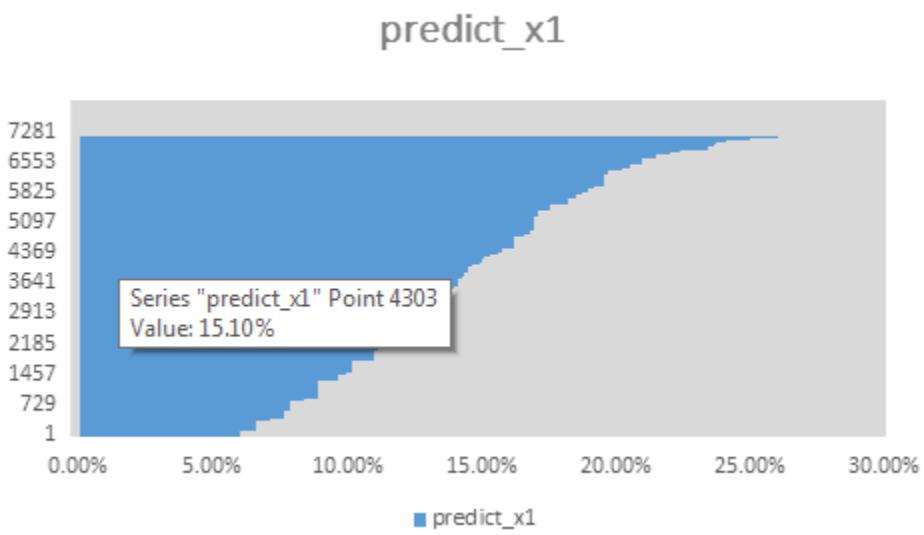
We successfully determined interest rate values for our “holdout dataset”.

Inferences

- Minimum value of interest rate is 9.99% and maximum value is 26.06%.
- Number of payments are either 36 or 60.
- Loans are divided in grades and subgrades.

Recommendations

- We have got much expected accuracy on prediction of interest rates.
- We have taken 8000 obs. Dataset and made a chart for its interest rate as shown below:



- It shows as that our interest lies on same range where maximum no. of values resides in middle part(10% to 20%) range.
- It gives us an overview of randomly selected 8000 obs. Interest rates via a chart.

References

- https://www.analyticsvidhya.com/blog/2016/05/h2o-data-table-build-models-large-data-sets/