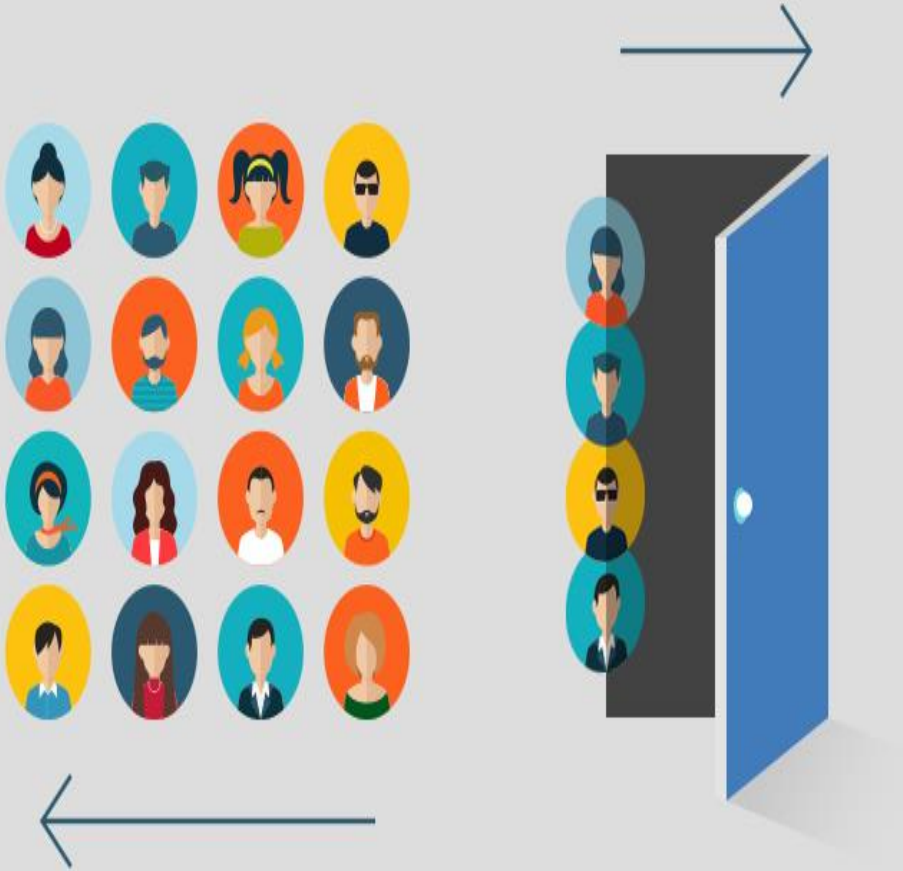
Identifying customers – with high probability of conversion



Data Types

ID	NAME	TYPE
A	ID	Character
B	Gender	Character
C	DOB	Numeric
D	Lead_Creation_Date	Numeric
E	City_Code	Character
F	City_Category	Character
G	Employer_Code	Character
H	Employer_Category1	Character
I	Employer_Category2	Numeric
J	Monthly_Income	Numeric
K	Customer_Existing_Primary_Bank_Code	Character

ID	NAME	TYPE
L	Primary_Bank_Type	Character
M	Contacted	Character
N	Source	Character
O	Source_Category	Character
P	Existing_EMI	Numeric
Q	Loan_Amount	Numeric
R	Loan_Period	Numeric
S	Interest_Rate	Numeric
T	EMI	Numeric
U	Var1	Numeric
V	Approved	Numeric

Data Summary

Type of Variable	Data Type	Variable Category	Observation
Predictor Variable - ID : Var1 - 21 Variables Target Variable - Approved	Character - A,B,E,F,G,H - K,L,M,N,O Numeric - C,D,I,J,P,Q,R,S,T,U - v - Approved	Categorical - A,B,E,F,G,H - K,L,M,N,O Continuous - C,D,I,J,P,Q - R,S,T,U	- 99751 Rows (Total) - 22 Variables Missing Values - Categorical - B,E,F,G,H,I,K,P - Continuous - P,Q,R,S,T

The following analysis helps to identify factor that highly helpful for acquiring the customers and the suggested model will help us to identify the percentage of conversion.

Data Pre-Processing

Treating Missing Values

Continuous Variables

- Missing values in variables DOB (after calculating age from Lead_Creation_date) are imputed by using Median of all existing values.
- Loan_amount, Loan_period, Interest_rate,Employee_category2 nearly ~40% rows are missing i.e. (27709/69713 – Training set & 11871/30038 – Test set) so treating missing values as separate level.

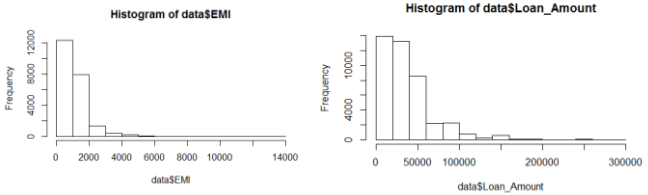
Categorical Variables

- Missing values in variables City_Code, City_Category, Employer_code, Employer_category_1, Customer_Existing_Primary_Bank_code, Primary_Bank_code, two type of method are tried to Impute
  - 1. imputing by using KNN imputation method.
  - 2. Treating missing values as separate level.

Data Pre-Processing

Normalization & Outlier Detection

- Transforming all continuous variables to Z-score scale, i.e. $z = \frac{x - \mu}{\sigma}$
- Outlier detected in Monthly_income, Existing_EMI
  - Monthly_income above 1,000,000 to 1,000,000
  - Exisitng_EMI above 500000 to its Median of its existing value (without NA and Zero)



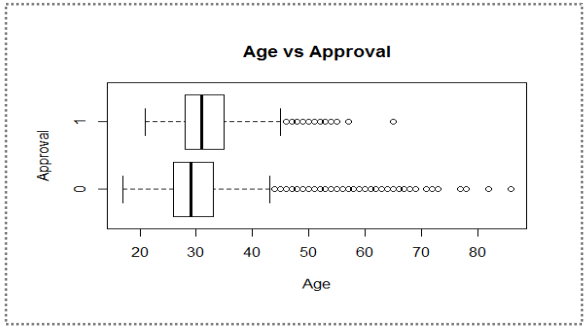
Histogram clearly states that, data is skewed to right indicating mean is greater than median.

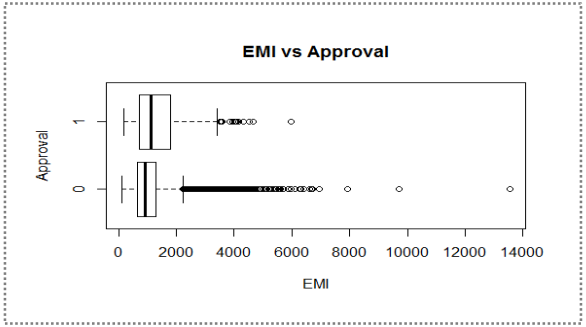
- In categorical variable City_code, Employer_code as too many levels, so grouping high number of city_code and labelling remaining as others similarly for Employer_code

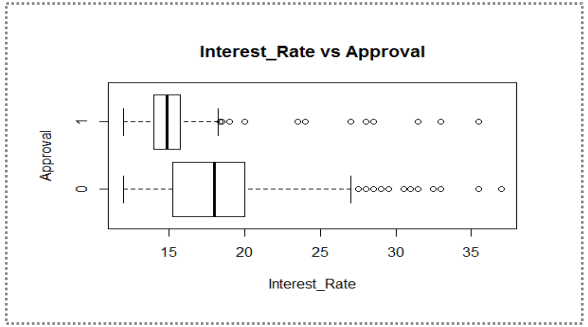
Variable Importance Analysis

Creating New Variables

- Age – from DOB – Lead_Creation_Date, because DOB has two many level to simplify that.
- Lead_Creation_Day – from Lead_Creation_Date, because month and year of Lead_Creation_Date has no level but day may bring some patter
- Net_Income – from Monthly_Income – Exisitng_EMI



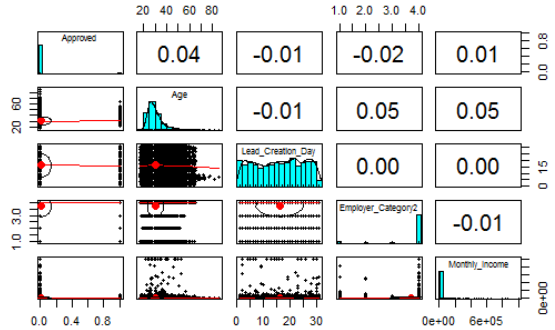


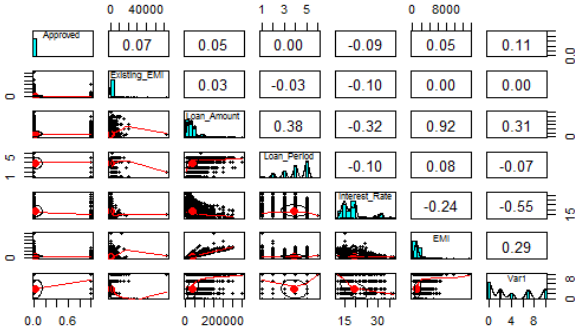


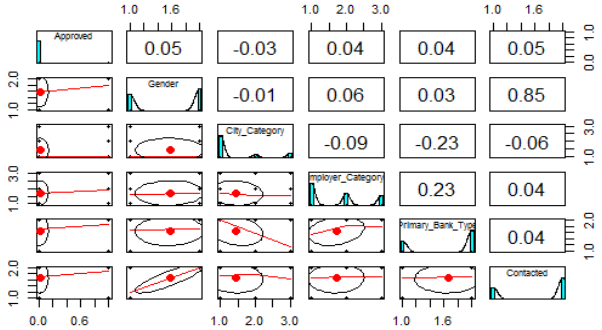
Approval rate is higher when Age has low values but still some outlier exist.

Approval status is distributed from mean of those having high Interest rate, those who are all not approved are skewed to right than those approved customers.

Variable Importance Analysis







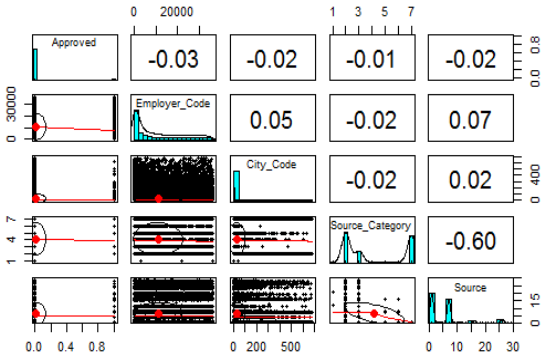


Chart in the left depict the variable correlation between predictor vs. target variable.

- Age, Var1, Existing_EMI, Interest_rate, Existing_EMI, Loan_Amount, Monthly_Income have high correlation with Renewal subscription.
- Source, Source_Category have low correlation with Renewal subscription when compared to other variables.

Variable Inflation Factor

- Detecting Multi-collinearity
- VIF of jth independent variable $\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$
- Removing the variable with highest VIF (>5)

Round #1
- Variable Contacted is removed

Round #2
-Variable Employer_Code, City_Code Source_Category,Source is removed

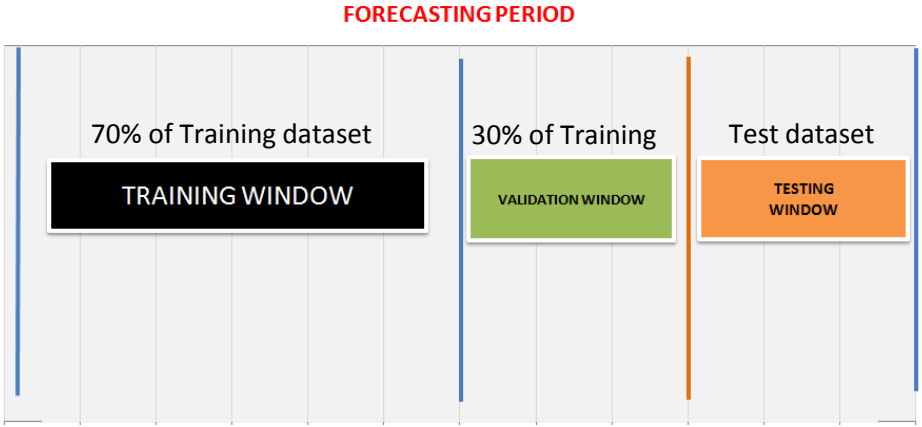
Considering above analysis, following variables are considered for modelling.

Categorical
- Gender, City_Category, Employer_Category1, Primary_Bank_Type

Continuous
- Age, Lead_Creation_Day, Employer_Category2, Monthly_Income
- Existing_EMI, Loan_Amount, Loan_Period, Interest_Rate, EMI, Var1

Pre-Modelling Summary

- 69713 – Training & 30038 – Test observation were considered. Considered Variable ,
  - Categorical
    - Gender, City_Category, Employer_Category1, Primary_Bank_Type
  - Continuous
    - Age, Lead_Creation_Day, Employer_Category2, Monthly_Income
    - Existing_EMI, Loan_Amount, Loan_Period, Interest_Rate, EMI, Var1
- Validation Set : Since we don't have Approved column in test set, to evaluate model splitting the training set into 70:30
- For deciding accuracy, Confusion Matrix : It creates confusion matrix and error images for each class. In addition it calculates the classification accuracy assessment indices (accuracy measure - adding true positive and true negative, divide by total).



Methodology

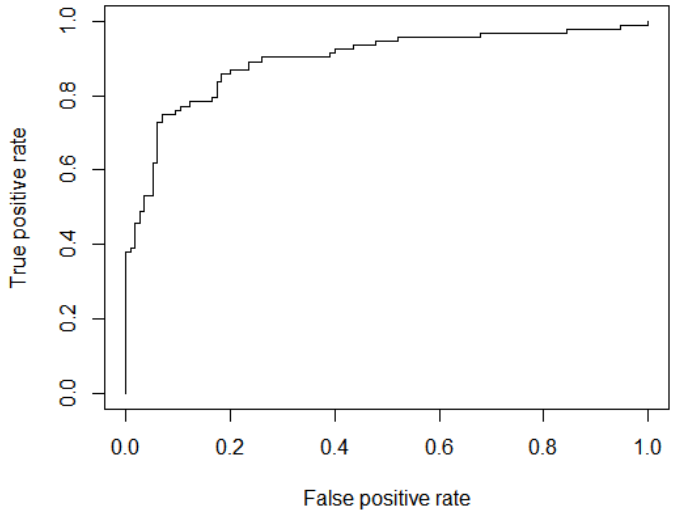
- Building model using 70% training of data and then using it to predict remaining 30% of training data (validation Period) to find out best working model.
- Best method is used to predict test dataset

AUC Measure at Validation Period

Period	Method	AUC
70% train dataset	Gradient Boosting Machine (GBM)	78%
70% train dataset	Random Forest	75%
70% train dataset	Logistic Regression	73%

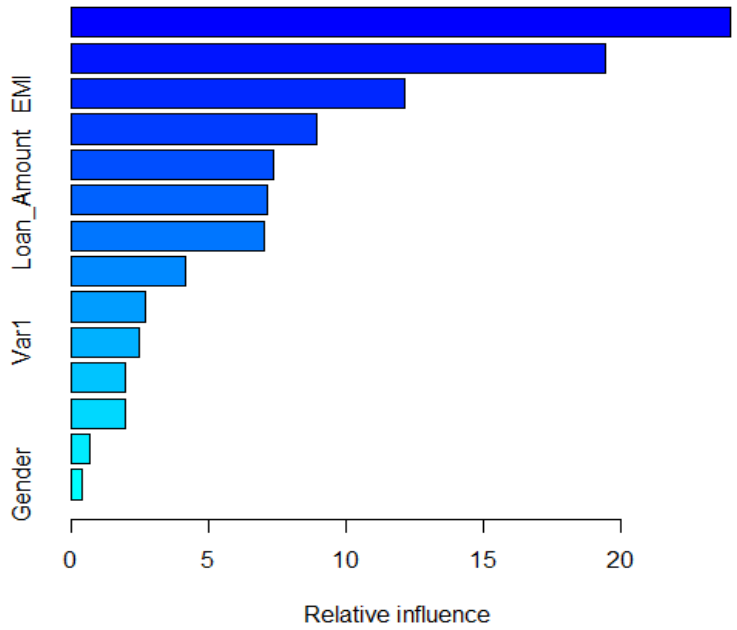
GBM method is used to predict the test data set.

ROC plot – validation Set



Variable Importance

Variable Importance Plot in GBM



NAME	TYPE
Monthly_Income	23.99511
Existing_EMI	19.44356
EMI	12.10537
Interest_Rate	8.936459
Age	7.353271
Loan_Amount	7.100448
Lead_Creation_Day	6.981453
Employer_Category1	4.110517
Primary_Bank_Type	2.696924
Var1	2.426097
Loan_Period	1.954906
Employer_Category2	1.934144
City_Category	0.62325
Gender	0.338495

Deliverables

- Variables Monthly_Income, Existing_EMI, EMI, Interest_Rate has significant impact in Approval status.
- variables like Age , Loan_Amount , Lead_Creation_Day, Employer_Category1 has some impact in Approval status.
- Potential customers are based on the following variables, Monthly_Income, Existing_EMI, EMI, Interest_Rate, Age , Loan_Amount , Lead_Creation_Day.