Santander Customer Transaction Prediction

Santander has provided with lakhs of anonymous data set which contain 200 numeric feature variable, target(0 & 1) & ID code. Our objective is to identify whether a customer will make future transaction or not.

Both train & test has 2 lakh rows with variable **var_0** to **var_199** which is of float data type. This is a binary classification problem under supervised machine learning where we have to predict whether the customer will make future transaction or not i.e(0 & 1).

Explotary data analysis

```
$ target : int  0 0 0 0 0 0 0 0 0 0 ...
 $ var_0  : num  8.93 11.5 8.61 11.06 9.84 ...
 $ var_1  : num  -6.79 -4.15 -2.75 -2.15 -1.48 ...
 $ var_2  : num  11.91 13.86 12.08 8.95 12.87 ...
 $ var_3  : num  5.09 5.39 7.89 7.2 6.64 ...
 $ var_4  : num  11.5 12.4 10.6 12.6 12.3 ...
 $ var_5  : num  -9.28 7.04 -9.08 -1.84 2.45 ...
 $ var_6  : num  5.12 5.62 6.94 5.84 5.94 ...
 $ var_7  : num  18.6 16.5 14.6 14.9 19.3 ...
```

Here all variables are of numeric data type. So we don’t required to change the data type.

Missing Value

Sample

```
    Missing_Percentage Columns

2                    0   var_0
3                    0   var_1
4                    0   var_2
5                    0   var_3
6                    0   var_4
7                    0   var_5
8                    0   var_6
9                    0   var_7
10                   0   var_8
11                   0   var_9
12                   0  var_10
13                   0  var_11
```

Missing value is one of the important factor we wave to check when we get the data, but in this Santander data we don’t have any NA or missing value. So we don’t have to drop or impute any variable.

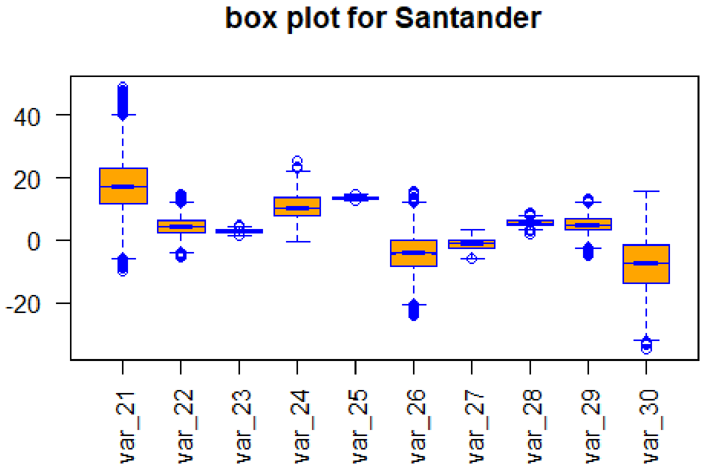
OUTLIERS

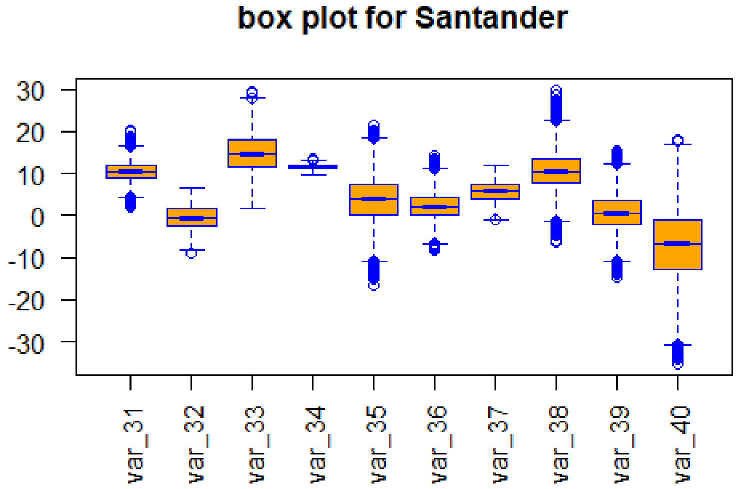
Outliers are the data which fall away from dataset.

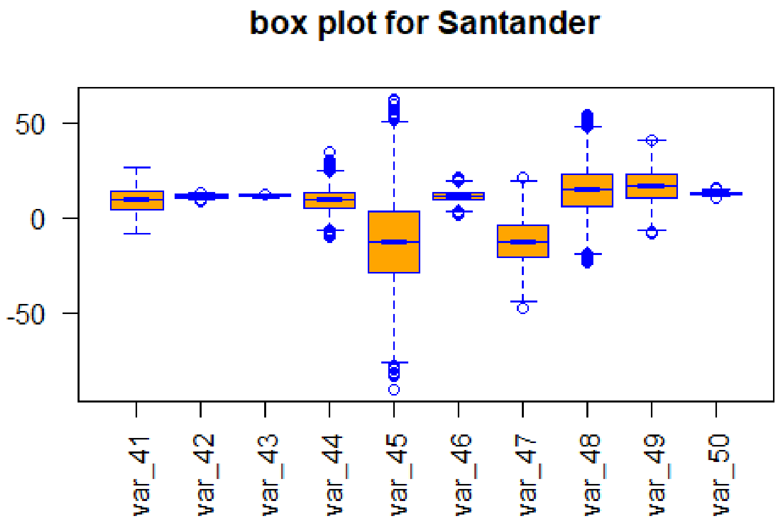
Causes of outliers

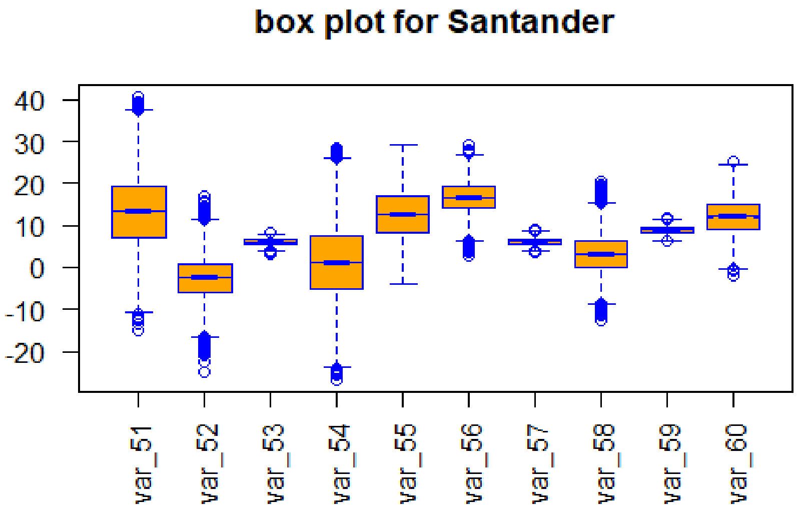
-Poor Data quality/contamination

-Lower quality measurement,malfunctioning equipatient etc





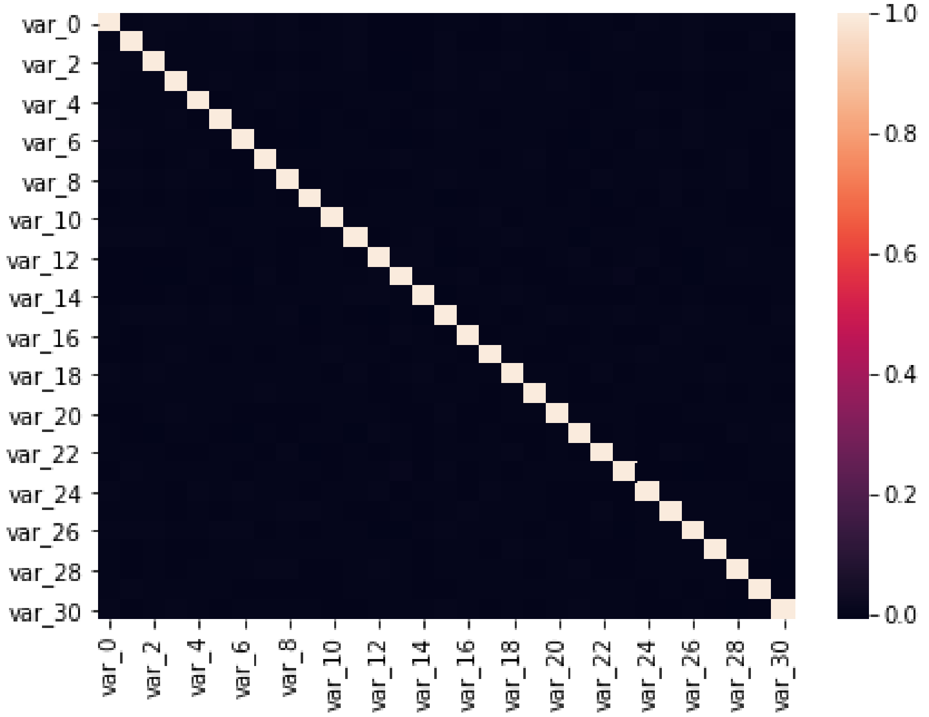




As in the sample box plot of 40 variables we can see some outliers,but as it is annonymous data we don’t know whether it is due to incorrect entered or measured data , so we will not remove it.

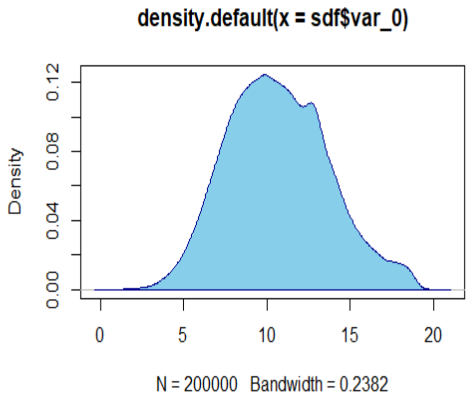
FEATURE SELECTION

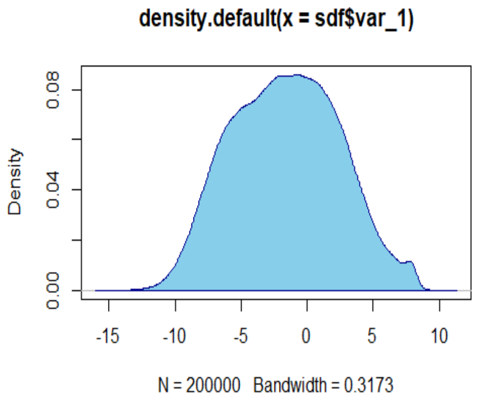
Also known as variable selection means selecting a subset of relevant feature for use in the model selection.

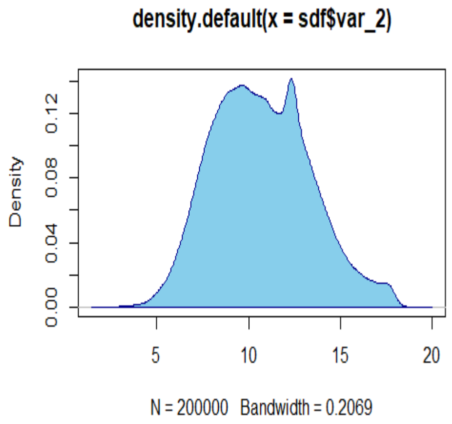


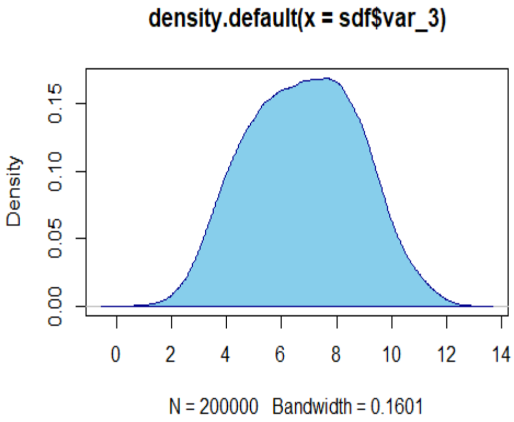
As a sample we have taken first 30 variable & check for their collinearity, In the above heat map we can clearly see no variable is correlated with each other. So, we can assume of sending all variable to the model.

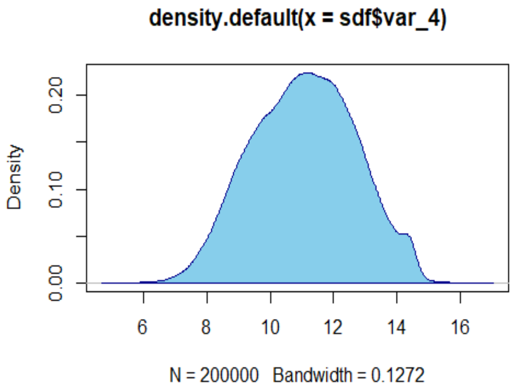
CHECKING NORMALIZATION

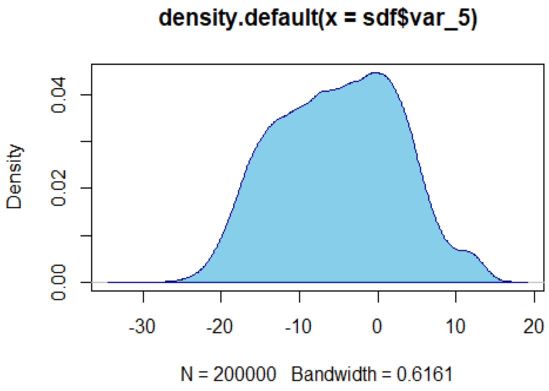


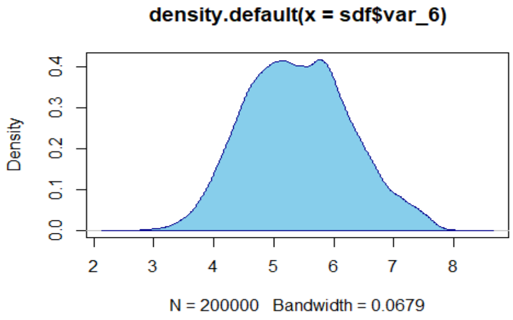


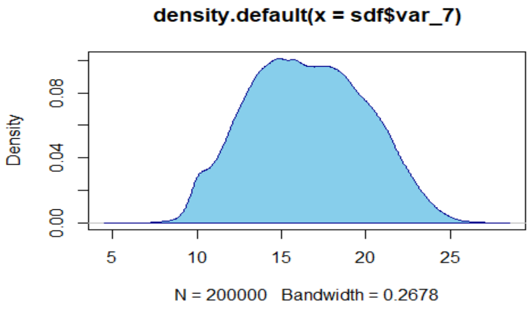

density.default(x = sdf$var_2)
Density
0.00
0.04
0.08
0.12
5
10
15
20
N = 200000 Bandwidth = 0.2069

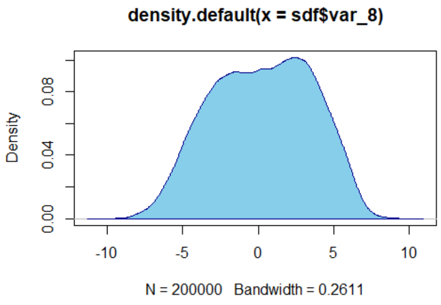

density.default(x = sdf$var_3)
Density
0.00
0.05
0.10
0.15
0
2
4
6
8
10
12
14
N = 200000 Bandwidth = 0.1601

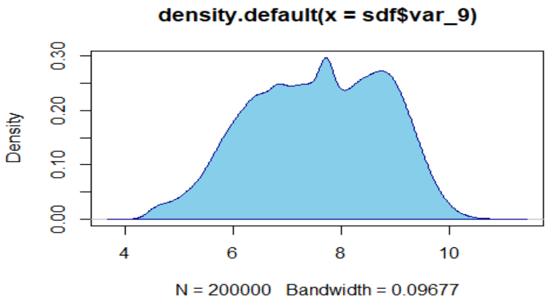

density.default(x = sdf$var_4)
Density
0.00
0.10
0.20
6
8
10
12
14
16
N = 200000 Bandwidth = 0.1272

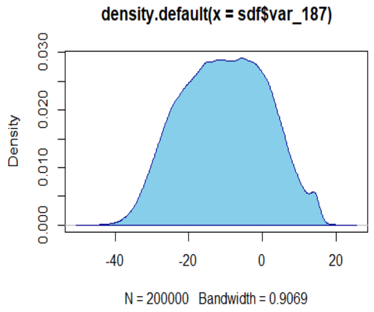

density.default(x = sdf$var_5)
Density
0.00
0.02
0.04
-30
-20
-10
0
10
20
N = 200000 Bandwidth = 0.6161

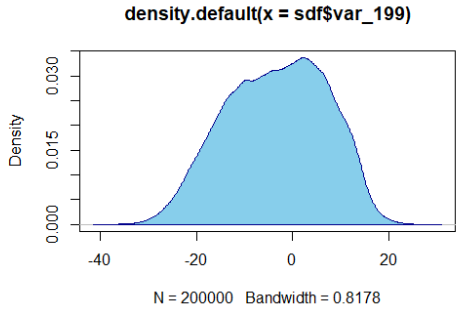

density.default(x = sdf$var_6)
Density
0.0
0.1
0.2
0.3
0.4
2
3
4
5
6
7
8
N = 200000 Bandwidth = 0.0679

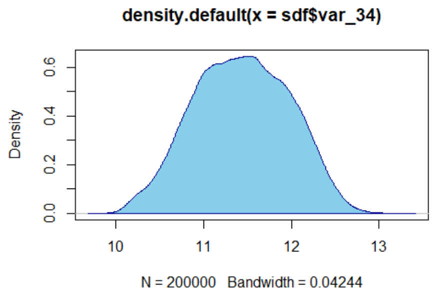

density.default(x = sdf$var_7)
Density
0.00
0.04
0.08
5
10
15
20
25
N = 200000 Bandwidth = 0.2678

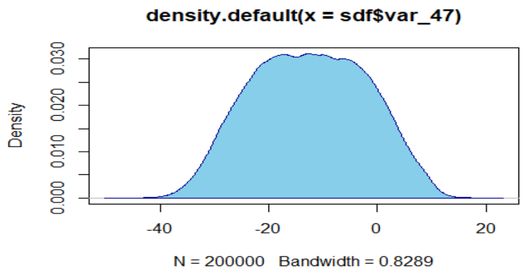

density.default(x = sdf$var_8)
Density
0.00
0.04
0.08
-10
-5
0
5
10
N = 200000 Bandwidth = 0.2611


density.default(x = sdf$var_9)
Density
0.00
0.10
0.20
0.30
4
6
8
10
N = 200000 Bandwidth = 0.09677









Looking at the above graph of some variable we can say that every variable in the data set are normally distributed.

LOGISTIC REGRESSION

- Logistic regression are used in classification model where the outcomes are in probabilities which can be used in binomial, ordinal & multinomial.

```
Call:
glm(formula = target ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6893  -0.3991  -0.2313  -0.1231   3.8072

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.279e+01  7.398e+00   8.488  < 2e-16 ***
var_0        5.650e-02  3.350e-03  16.867  < 2e-16 ***
var_1        4.016e-02  2.549e-03  15.755  < 2e-16 ***
var_2        6.304e-02  3.855e-03  16.354  < 2e-16 ***
var_3        1.804e-02  5.078e-03   3.553 0.000381 ***
var_4        2.443e-02  6.359e-03   3.842 0.000122 ***
var_5        1.290e-02  1.312e-03   9.829  < 2e-16 ***
var_6        2.716e-01  1.181e-02  22.996  < 2e-16 ***
```

```
var_7       -8.835e-04  3.026e-03  -0.292 0.770329    
var_8        1.761e-02  3.117e-03   5.650 1.61e-08 ***
var_9       -1.035e-01  8.315e-03 -12.443  < 2e-16 ***
var_10      -5.411e-04  1.886e-03  -0.287 0.774124    
var_11       1.326e-02  1.727e-03   7.679 1.61e-14 ***
var_12      -1.150e+00  5.356e-02 -21.479  < 2e-16 ***
var_13      -4.119e-02  2.212e-03 -18.623  < 2e-16 ***
var_14      -5.559e-03  4.617e-03  -1.204 0.228506    
var_15       1.325e-01  2.515e-02   5.267 1.38e-07 ***
var_16       1.121e-02  4.041e-03   2.775 0.005519 ** 
var_17       7.415e-04  1.548e-03   0.479 0.631843    
var_18       1.741e-02  1.308e-03  13.310  < 2e-16 ***
var_19       4.174e-03  1.293e-03   3.228 0.001247 ** 
var_20      -9.404e-03  1.765e-03  -5.329 9.88e-08 ***
var_21      -2.341e-02  1.261e-03 -18.567  < 2e-16 ***
var_22       7.121e-02  3.591e-03  19.829  < 2e-16 ***
var_23      -1.755e-01  1.967e-02  -8.922  < 2e-16 ***
var_24       2.846e-02  2.729e-03  10.427  < 2e-16 ***
var_25       1.337e-01  3.619e-02   3.694 0.000221 ***
var_26       3.498e-02  1.720e-03  20.344  < 2e-16 ***
var_27      -6.568e-03  6.806e-03  -0.965 0.334504    
var_28      -1.148e-01  1.322e-02  -8.684  < 2e-16 ***
var_29       5.829e-03  3.963e-03   1.471 0.141335    
var_30       5.887e-04  1.306e-03   0.451 0.652285    
var_31      -4.119e-02  4.813e-03  -8.558  < 2e-16 ***
var_32       3.842e-02  3.999e-03   9.608  < 2e-16 ***
var_33      -3.498e-02  2.406e-03 -14.541  < 2e-16 ***
var_34      -3.157e-01  1.911e-02 -16.521  < 2e-16 ***
var_35       2.319e-02  1.995e-03  11.628  < 2e-16 ***
var_36      -3.801e-02  3.318e-03 -11.455  < 2e-16 ***
var_37       1.312e-02  4.590e-03   2.859 0.004252 ** 
var_38       1.224e-03  2.426e-03   0.505 0.613790    
var_39      -3.229e-03  2.547e-03  -1.268 0.204854    
var_40       2.103e-02  1.243e-03  16.927  < 2e-16 ***
var_41      -1.154e-05  1.748e-03  -0.007 0.994736    
var_42      -3.591e-02  1.486e-02  -2.416 0.015690 *  
var_43      -2.717e-01  3.336e-02  -8.145 3.80e-16 ***
var_44      -2.736e-02  1.713e-03 -15.974  < 2e-16 ***
var_45      -3.327e-03  4.847e-04  -6.865 6.67e-12 ***
var_46       7.210e-03  3.611e-03   1.997 0.045879 *  
var_47       2.802e-03  9.813e-04   2.855 0.004298 ** 
var_48       8.628e-03  9.108e-04   9.473  < 2e-16 ***
var_49       1.245e-02  1.319e-03   9.443  < 2e-16 ***
var_50      -5.840e-02  1.490e-02  -3.920 8.84e-05 ***
var_51       8.772e-03  1.258e-03   6.971 3.14e-12 ***
var_52       1.980e-02  2.081e-03   9.513  < 2e-16 ***
var_53       2.849e-01  1.347e-02  21.148  < 2e-16 ***
var_54      -7.724e-03  1.232e-03  -6.267 3.68e-10 ***
var_55       9.874e-03  1.819e-03   5.428 5.71e-08 ***
var_56      -3.275e-02  2.898e-03 -11.300  < 2e-16 ***
var_57      -7.769e-02  1.303e-02  -5.964 2.47e-09 ***
var_58      -2.054e-02  2.401e-03  -8.556  < 2e-16 ***
var_59      -4.930e-02  1.212e-02  -4.067 4.77e-05 ***
var_60       6.298e-03  2.448e-03   2.572 0.010102 *  
var_61       2.073e-03  8.924e-04   2.323 0.020156 *  
var_62       2.799e-02  5.078e-03   5.511 3.57e-08 ***
var_63      -1.681e-02  3.317e-03  -5.068 4.03e-07 ***
```

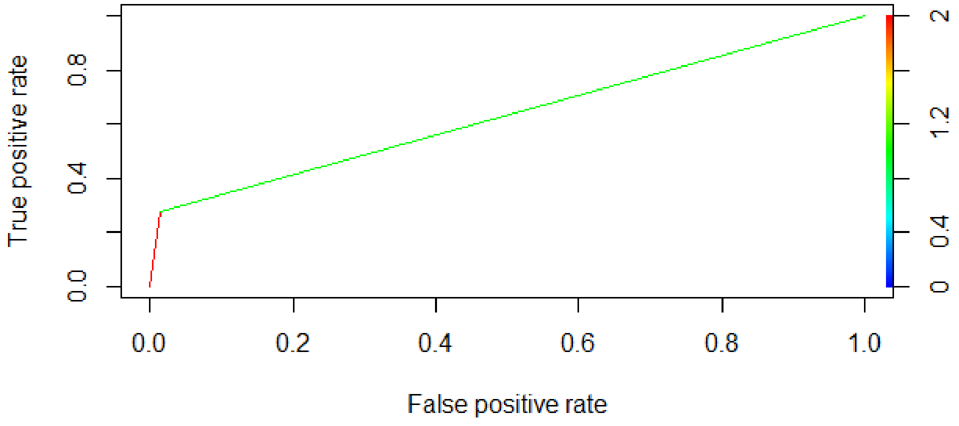
```
var_64      -2.798e-02  6.949e-03  -4.027 5.66e-05 ***
var_65       8.649e-03  2.755e-03   3.140 0.001690 ** 
var_66       6.380e-02  9.187e-03   6.944 3.80e-12 ***
var_67       1.917e-02  1.408e-03  13.616  < 2e-16 ***
var_68      -5.626e+00  1.438e+00  -3.912 9.14e-05 ***
var_69       7.238e-03  2.610e-03   2.773 0.005551 ** 
var_70       7.923e-03  8.683e-04   9.125  < 2e-16 ***
var_71       4.069e-01  3.881e-02  10.485  < 2e-16 ***
var_72      -1.003e-02  2.617e-03  -3.831 0.000128 ***
var_73      -3.371e-03  1.388e-03  -2.428 0.015181 *  
var_74       4.766e-03  7.363e-04   6.472 9.67e-11 ***
var_75      -2.080e-02  1.697e-03 -12.254  < 2e-16 ***
var_76      -2.540e-02  1.291e-03 -19.674  < 2e-16 ***
var_77      -1.544e-02  2.727e-03  -5.663 1.49e-08 ***
var_78       7.840e-02  5.196e-03  15.089  < 2e-16 ***
var_79       8.570e-03  7.885e-03   1.087 0.277085    
var_80      -2.593e-02  1.367e-03 -18.972  < 2e-16 ***
var_81      -1.120e-01  4.393e-03 -25.497  < 2e-16 ***
var_82       7.378e-03  1.219e-03   6.050 1.44e-09 ***
var_83      -7.642e-03  1.241e-03  -6.160 7.27e-10 ***
var_84       6.242e-03  1.666e-03   3.748 0.000178 ***
var_85      -1.756e-02  2.656e-03  -6.612 3.78e-11 ***
var_86      -1.709e-02  1.317e-03 -12.976  < 2e-16 ***
var_87      -2.131e-02  1.833e-03 -11.627  < 2e-16 ***
var_88      -2.429e-02  4.161e-03  -5.838 5.27e-09 ***
var_89       3.708e-02  2.884e-03  12.858  < 2e-16 ***
var_90       7.171e-03  7.898e-04   9.079  < 2e-16 ***
var_91       8.253e-01  6.744e-02  12.237  < 2e-16 ***
var_92      -3.568e-02  2.471e-03 -14.438  < 2e-16 ***
var_93      -2.152e-01  1.883e-02 -11.424  < 2e-16 ***
var_94       5.832e-02  3.718e-03  15.687  < 2e-16 ***
var_95       1.820e-01  1.652e-02  11.014  < 2e-16 ***
var_96       2.323e-03  1.212e-03   1.916 0.055348 .  
var_97       3.570e-03  8.193e-04   4.357 1.32e-05 ***
var_98      -7.280e-03  1.447e-02  -0.503 0.615013    
var_99       1.004e-01  5.494e-03  18.274  < 2e-16 ***
var_100      1.071e-03  1.132e-03   0.946 0.343908    
var_101     -7.569e-03  2.089e-03  -3.624 0.000290 ***
var_102     -6.948e-03  1.199e-03  -5.793 6.92e-09 ***
var_103     -5.723e-02  5.581e-02  -1.026 0.305097    
var_104     -5.097e-02  5.281e-03  -9.652  < 2e-16 ***
var_105      9.719e-02  1.207e-02   8.049 8.33e-16 ***
var_106      5.917e-02  5.455e-03  10.847  < 2e-16 ***
var_107     -1.921e-02  1.367e-03 -14.049  < 2e-16 ***
var_108     -8.817e-01  5.984e-02 -14.734  < 2e-16 ***
var_109     -3.716e-02  2.367e-03 -15.702  < 2e-16 ***
var_110      5.574e-02  2.672e-03  20.860  < 2e-16 ***
var_111      7.540e-02  9.481e-03   7.952 1.83e-15 ***
var_112      4.927e-02  6.544e-03   7.530 5.09e-14 ***
var_113     -1.174e-02  2.317e-03  -5.065 4.09e-07 ***
var_114     -9.565e-02  1.048e-02  -9.128  < 2e-16 ***
var_115     -5.724e-02  3.920e-03 -14.604  < 2e-16 ***
var_116     -5.333e-02  6.249e-03  -8.534  < 2e-16 ***
var_117      8.728e-04  7.743e-04   1.127 0.259668    
var_118      1.589e-02  1.178e-03  13.485  < 2e-16 ***
var_119      2.421e-02  2.457e-03   9.853  < 2e-16 ***
var_120     -2.731e-03  8.542e-04  -3.197 0.001387 ** 
```

```
var_121     -8.176e-02  6.073e-03 -13.462  < 2e-16 ***
var_122     -2.771e-02  1.997e-03 -13.875  < 2e-16 ***
var_123     -2.112e-02  1.662e-03 -12.705  < 2e-16 ***
var_124      7.051e-03  3.786e-03   1.863 0.062500 .
var_125      3.162e-01  3.227e-02   9.799  < 2e-16 ***
var_126      9.275e-03  1.334e-02   0.696 0.486736
var_127     -3.808e-02  3.284e-03 -11.595  < 2e-16 ***
var_128      2.818e-02  3.188e-03   8.839  < 2e-16 ***
var_129     -6.053e-03  2.504e-03  -2.418 0.015616 *
var_130      1.207e-01  1.242e-02   9.721  < 2e-16 ***
var_131     -1.732e-01  2.255e-02  -7.682 1.57e-14 ***
var_132     -6.168e-02  7.085e-03  -8.707  < 2e-16 ***
var_133      4.631e-01  2.722e-02  17.009  < 2e-16 ***
var_134      9.407e-03  1.675e-03   5.614 1.97e-08 ***
var_135      1.259e-02  1.355e-03   9.295  < 2e-16 ***
var_136     -1.008e-03  9.983e-04  -1.010 0.312432
var_137      1.123e-02  1.166e-03   9.625  < 2e-16 ***
var_138      1.198e-02  2.282e-03   5.252 1.51e-07 ***
var_139     -3.103e-02  1.329e-03 -23.353  < 2e-16 ***
var_140      1.216e-02  2.120e-03   5.737 9.65e-09 ***
var_141     -1.430e-02  1.537e-03  -9.302  < 2e-16 ***
var_142     -1.109e-02  1.816e-03  -6.105 1.03e-09 ***
var_143     -1.522e-02  3.523e-03  -4.321 1.56e-05 ***
var_144      8.165e-02  1.121e-02   7.286 3.19e-13 ***
var_145      2.489e-02  2.650e-03   9.393  < 2e-16 ***
var_146     -8.184e-02  4.046e-03 -20.230  < 2e-16 ***
var_147      1.823e-02  1.394e-03  13.078  < 2e-16 ***
var_148     -8.629e-01  5.146e-02 -16.769  < 2e-16 ***
var_149     -1.424e-02  9.969e-04 -14.284  < 2e-16 ***
var_150     -3.925e-02  4.203e-03  -9.338  < 2e-16 ***
var_151      2.325e-02  2.592e-03   8.968  < 2e-16 ***
var_152     -1.283e-02  3.435e-03  -3.735 0.000188 ***
var_153     -9.912e-03  5.157e-03  -1.922 0.054631 .
var_154     -2.836e-02  2.074e-03 -13.679  < 2e-16 ***
var_155      2.058e-02  1.781e-03  11.552  < 2e-16 ***
var_156     -7.207e-02  1.084e-02  -6.647 2.99e-11 ***
var_157      2.003e-02  1.843e-03  10.865  < 2e-16 ***
var_158     -2.807e-03  1.315e-03  -2.134 0.032808 *
var_159      1.232e-02  2.512e-03   4.902 9.47e-07 ***
var_160     -8.968e-04  9.511e-04  -0.943 0.345710
var_161      6.314e-02  4.754e-02   1.328 0.184140
var_162      7.363e-02  7.267e-03  10.133  < 2e-16 ***
var_163      1.988e-02  1.948e-03  10.206  < 2e-16 ***
var_164      2.665e-02  1.903e-03  14.008  < 2e-16 ***
var_165     -3.590e-02  2.057e-03 -17.456  < 2e-16 ***
var_166     -4.980e-01  2.784e-02 -17.888  < 2e-16 ***
var_167      1.131e-02  1.324e-03   8.547  < 2e-16 ***
var_168      1.082e-02  3.304e-03   3.276 0.001053 **
var_169     -4.077e-01  2.796e-02 -14.580  < 2e-16 ***
var_170      3.688e-02  2.320e-03  15.899  < 2e-16 ***
var_171      9.674e-03  1.919e-03   5.042 4.60e-07 ***
var_172     -1.505e-02  1.195e-03 -12.598  < 2e-16 ***
var_173      2.420e-02  1.738e-03  13.921  < 2e-16 ***
var_174     -2.888e-02  1.436e-03 -20.110  < 2e-16 ***
var_175      2.943e-02  3.562e-03   8.263  < 2e-16 ***
var_176      4.357e-03  1.381e-03   3.154 0.001609 **
var_177     -4.953e-02  3.949e-03 -12.542  < 2e-16 ***
```

```
var_178     -6.860e-03  1.204e-03  -5.697 1.22e-08 ***
var_179      5.314e-02  3.625e-03  14.657  < 2e-16 ***
var_180      2.008e-02  1.958e-03  10.256  < 2e-16 ***
var_181      3.321e-02  7.536e-03   4.407 1.05e-05 ***
var_182     -3.192e-03  1.158e-03  -2.757 0.005826 **
var_183     -2.053e-03  2.317e-03  -0.886 0.375498
var_184      1.749e-02  1.105e-03  15.824  < 2e-16 ***
var_185     -4.060e-05  2.194e-03  -0.019 0.985236
var_186     -2.794e-02  3.246e-03  -8.607  < 2e-16 ***
var_187      4.640e-03  8.972e-04   5.172 2.31e-07 ***
var_188     -3.019e-02  2.625e-03 -11.500  < 2e-16 ***
var_189      3.109e-02  1.062e-02   2.927 0.003418 **
var_190      3.971e-02  2.269e-03  17.500  < 2e-16 ***
var_191      5.207e-02  3.375e-03  15.429  < 2e-16 ***
var_192     -9.709e-02  7.052e-03 -13.769  < 2e-16 ***
var_193     -1.568e-02  2.597e-03  -6.037 1.57e-09 ***
var_194     -2.381e-02  3.299e-03  -7.216 5.37e-13 ***
var_195      6.436e-02  7.191e-03   8.950  < 2e-16 ***
var_196      1.266e-02  1.896e-03   6.680 2.39e-11 ***
var_197     -1.399e-01  1.122e-02 -12.475  < 2e-16 ***
var_198     -5.953e-02  3.389e-03 -17.564  < 2e-16 ***
 [ reached getOption("max.print") -- omitted 1 row ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 91326  on 139999  degrees of freedom
Residual deviance: 64553  on 139799  degrees of freedom
AIC: 64955

Number of Fisher Scoring iterations: 6
```

-Min -2.68 & Max 3.80 means that there is no much deviation in error.
- Pr(>|z|) says the amount of significance of each variable in deciding the target variable, the star mark after the value tells that how strong that variable is in deciding the target.
-Null Deviance will tell how well the target variable is predicted by the model with the help of intercept.
-Residual Deviance tell how well the target variable is predicated using Null Deviance & other independent variable.
-AIC(Akaike information criterion) will help to choose better model,less AIC give more accuracy.

ROC Curve



ROC curve is performance measurement for classification problem for various threshold settings. ROC is a probability curve & AUC represents degree or measure of separibility.Higher the AUC ,better is the model in predicting 0's & 1's.

ROC curve is plotted with True Positive rate & False Positive Rate with False Positive rate in x-axis & True Positive Rate in Y-axis.

Precision & Recall, Precision means percentage of results which are relevant,recall refers to the percentage of total relevant results correctly classified by your algorithm

Error Metrics for Logistic regression

R Code

TP=53204
FP=4368
FN=767
TN=1661

Recall=98.5

Accuracy=91.4

Precision=92.4

AUC= 63.0

Python Code

TP=35529
FP=2942
FN=494
TN=1128

Recall=27.71

Accuracy=91.42

Precision=69.98

DECISION TREE

A predictive model based on a branching series of boolean test
-can be used for classification & regression

2 popular decision tree algorithm'

C5.0,CART

Error Metrics for Decision Tree

R Code

TP=53181
TN=472
FN=789
FP=5557

Recall=98.5

Accuracy=89.4

Precision=90.53

AUC= 53.18

Python Code

TP=49008
FP=4898
FN=4956
TN=1138

Recall=18.85

Accuracy=83.57

Precision=18.67

NAÏVE BAYES

-Naïve Bayes is a Probabilistic Classification Algorithm
-It works on Bayes Theorem of probability to predict the class of unknown dataset

Error Metrics for Decision Tree

R Code

TP=53072
TN=2212
FN=899
FP=3817

Recall=98.33

Accuracy=92.14

Precision=93.29

AUC= 67.51

Python Code

TP=53099
FP=3812
FN=865
TN=2224

Recall=36.84

Accuracy=92.20

Precision=71.99

Conclusion: Accuracy is Naïve Bayes is more. Therefore we will select this model for prediction.