Project

Churn reduction

Sankepally Vikram Reddy

19/05/2018

Contents

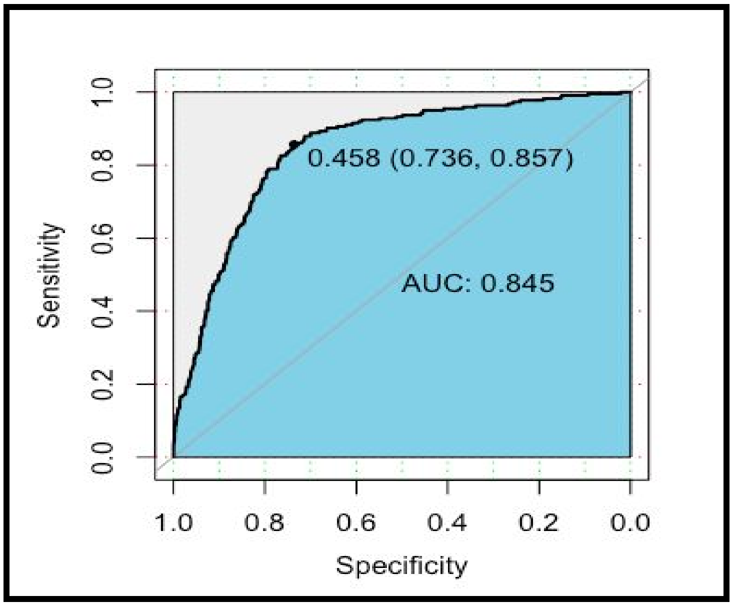
Chapter 1

R code sample output

Logistic regression(with outliers)

```
Generalized Linear Model

3333 samples
  10 predictor
   2 classes: 'False', 'True'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3000, 2999, 3000, 3000, 3000, 2999, ...
Resampling results:

  ROC        Sens      Spec
  0.7977286  0.776882  0.699076



Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6715  -0.8868  -0.4113   0.9171   2.6147

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -4.7963996  0.3627849 -13.221  < 2e-16 ***
international.plan2          2.1901153  0.1291722  16.955  < 2e-16 ***
voice.mail.plan2            -0.6961440  0.0986332  -7.058 1.69e-12 ***
total.day.minutes            0.0102119  0.0006457  15.816  < 2e-16 ***
total.eve.minutes            0.0018952  0.0011596   1.634  0.10221
total.eve.charge             0.0387134  0.0136926   2.827  0.00469 **
total.night.minutes          0.0019929  0.0007444   2.677  0.00742 **
total.night.calls            0.0005099  0.0018297   0.279  0.78048
total.intl.minutes           0.0420451  0.0128113   3.282  0.00103 **
total.intl.calls            -0.0363787  0.0141111  -2.578  0.00994 **
number.customer.service.calls  0.4454228  0.0253844  17.547  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4617.0  on 3332  degrees of freedom
Residual deviance: 3656.7  on 3322  degrees of freedom
AIC: 3678.7

Number of Fisher Scoring iterations: 4


pred_class  False.  True.
     False    1123     47
     True      320    177
[1] "threshold"
[1] 0.4836551
Confusion Matrix and Statistics

          Reference
Prediction  False.  True.
    False.    1104    339
    True.       39    185

               Accuracy : 0.7732
                 95% CI : (0.7524, 0.7932)
    No Information Rate : 0.6857
    P-Value [Acc > NIR] : 1.34e-15

                  Kappa : 0.3774
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9659
            Specificity : 0.3531
         Pos Pred Value : 0.7651
         Neg Pred Value : 0.8259
             Prevalence : 0.6857
         Detection Rate : 0.6623
   Detection Prevalence : 0.8656
      Balanced Accuracy : 0.6595

       'Positive' Class :  False.
```

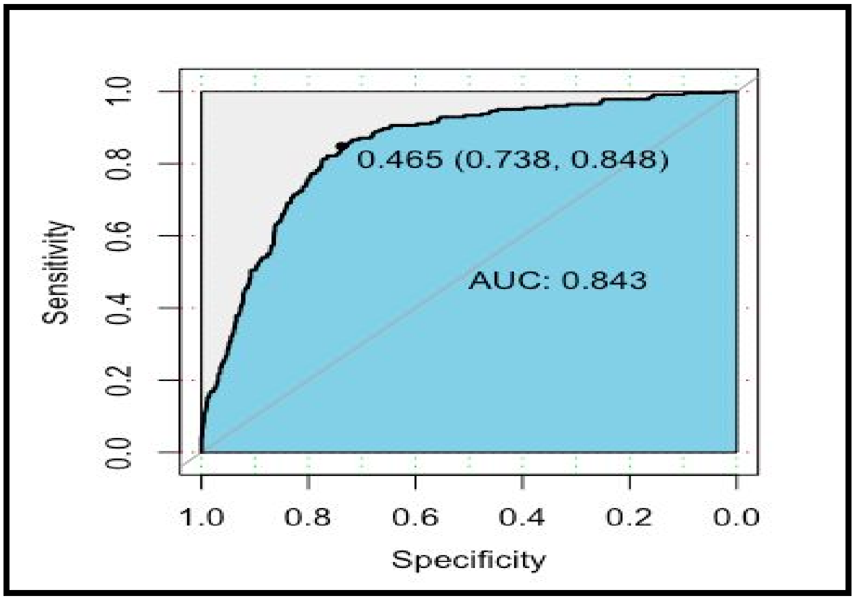


Logistic regression (Without outliers)

```
Generalized Linear Model

3333 samples
  10 predictor
   2 classes: 'False', 'True'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3000, 3000, 3000, 2999, 3000, 3000, ...
Resampling results:

  ROC        Sens       Spec
  0.7932534  0.7716461  0.6873208



Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6870  -0.8955  -0.4291   0.9339   2.6155
```

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -4.9110187  0.3898845 -12.596  < 2e-16 ***
international.plan2        2.1614419  0.1281304  16.869  < 2e-16 ***
voice.mail.plan2          -0.6931057  0.0982341  -7.056 1.72e-12 ***
total.day.minutes          0.0107445  0.0007122  15.086  < 2e-16 ***
total.eve.minutes          0.0017485  0.0012724   1.374  0.16938
total.eve.charge           0.0417870  0.0150165   2.783  0.00539 **
total.night.minutes        0.0023554  0.0008035   2.931  0.00337 **
total.night.calls          0.0010189  0.0019809   0.514  0.60698
total.intl.minutes         0.0402223  0.0141160   2.849  0.00438 **
total.intl.calls          -0.0557369  0.0164249  -3.393  0.00069 ***
number.customer.service.calls  0.4368107  0.0251486  17.369  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4617.0  on 3332  degrees of freedom
Residual deviance: 3689.6  on 3322  degrees of freedom
AIC: 3711.6

Number of Fisher Scoring iterations: 4


pred_class  False.  True.
     False    1119    46
     True      324   178
[1] "threshold"
[1] 0.4955664
Confusion Matrix and Statistics

          Reference
Prediction  False.  True.
    False.    1116    327
    True.       42    182

               Accuracy : 0.7786
                 95% CI : (0.7579, 0.7984)
```

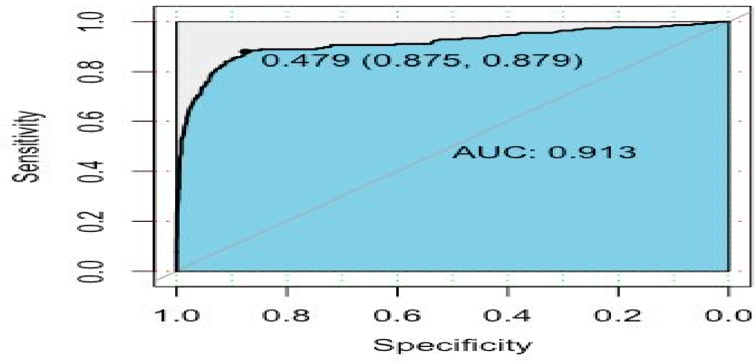
```
         No Information Rate : 0.6947
         P-Value [Acc > NIR] : 1.065e-14

                       Kappa : 0.3811
      Mcnemar's Test P-Value : < 2.2e-16

                 Sensitivity : 0.9637
                 Specificity : 0.3576
              Pos Pred Value : 0.7734
              Neg Pred Value : 0.8125
                  Prevalence : 0.6947
              Detection Rate : 0.6695
        Detection Prevalence : 0.8656
           Balanced Accuracy : 0.6606

            'Positive' Class :  False
```



Random forest (with outliers)

```
Random Forest

3333 samples
  10 predictor
   2 classes: 'False', 'True'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2999, 3000, 3000, 2999, 3000, 3000, ...
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
   2    0.8833554  0.8402272  0.7778928
   6    0.8815604  0.8408086  0.7816042
  10    0.8776084  0.8349913  0.7846983

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
                Length Class      Mode
call               6   -none-     call
type               1   -none-     character
predicted       3333   factor     numeric
err.rate        1500   -none-     numeric
confusion          6   -none-     numeric
votes           6666   matrix     numeric
oob.times       3333   -none-     numeric
classes            2   -none-     character
importance        10   -none-     numeric
importanceSD       0   -none-     NULL
localImportance    0   -none-     NULL
proximity          0   -none-     NULL
ntree              1   -none-     numeric
mtry               1   -none-     numeric
forest            14   -none-     list
y               3333   factor     numeric
test               0   -none-     NULL
inbag              0   -none-     NULL
xNames            10   -none-     character
problemType        1   -none-     character
tuneValue          1   data.frame list
obsLevels          2   -none-     character
```

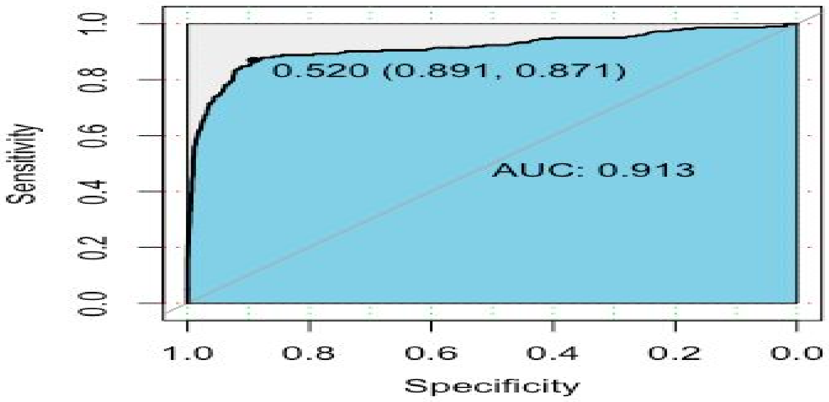
```
param             2   -none-     list

pred_class  False.  True.
    False    1272     31
    True      171    193
[1] "threshold"
[1] 0.479
Confusion Matrix and Statistics

          Reference
Prediction  False.  True.
   False.    1263    180
   True.       27    197

              Accuracy : 0.8758
                95% CI : (0.859, 0.8913)
   No Information Rate : 0.7738
   P-Value [Acc > NIR] : < 2.2e-16

                 Kappa : 0.5857
 Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.9791
           Specificity : 0.5225
        Pos Pred Value : 0.8753
        Neg Pred Value : 0.8795
            Prevalence : 0.7738
        Detection Rate : 0.7576
  Detection Prevalence : 0.8656
     Balanced Accuracy : 0.7508

       'Positive' Class :  False.
```



Random forest:Without outliers

```
Random Forest

3333 samples
  10 predictor
   2 classes: 'False', 'True'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3000, 3000, 2999, 2998, 3000, 3000, ...
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
   2    0.8834830  0.8489279  0.7822560
   6    0.8795385  0.8448582  0.7834944
  10    0.8769020  0.8343964  0.7872096

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
                Length Class  Mode
call               6   -none- call
type               1   -none- character
predicted       3333   factor numeric
err.rate        1500   -none- numeric
confusion          6   -none- numeric
votes           6666   matrix numeric
oob.times       3333   -none- numeric
classes            2   -none- character
importance        10   -none- numeric
importanceSD       0   -none- NULL
localImportance    0   -none- NULL
proximity          0   -none- NULL
ntree              1   -none- numeric
mtry               1   -none- numeric
forest            14   -none- list
y               3333   factor numeric
test               0   -none- NULL
inbag              0   -none- NULL
```

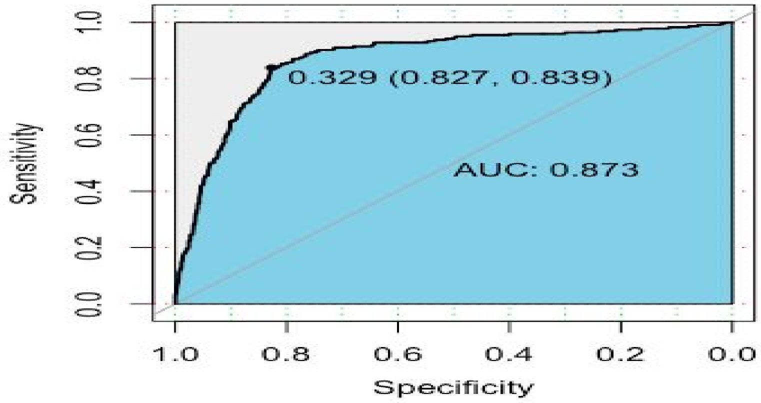
```
xNames           10   -none-     character
problemType       1   -none-     character
tuneValue         1   data.frame list
obsLevels         2   -none-     character
param             2   -none-     list

pred_class  False.  True.
    False    1270     28
    True      173    196
[1] "threshold"
[1] 0.52
Confusion Matrix and Statistics

          Reference
Prediction  False.  True.
    False.    1286    157
    True.       29    195

               Accuracy : 0.8884
                 95% CI : (0.8723, 0.9031)
    No Information Rate : 0.7888
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6136
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9779
            Specificity : 0.5540
         Pos Pred Value : 0.8912
         Neg Pred Value : 0.8705
             Prevalence : 0.7888
         Detection Rate : 0.7714
   Detection Prevalence : 0.8656
      Balanced Accuracy : 0.7660

       'Positive' Class :  False.
```



Naive bayes (with outliers)

```
With outliers
Naive Bayes

3333 samples
  10 predictor
   2 classes: 'False', 'True'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3000, 3000, 2999, 2999, 2999, 3000, ...
Resampling results across tuning parameters:

  fL  adjust  ROC        Sens       Spec
  0   0.0     0.0000000        NaN        NaN
  1   0.5     0.8309535  0.8797184  0.6097999
  3   2.0     0.8469729  0.9250403  0.5533471

Tuning parameter 'usekernel' was held constant at a value of TRUE
ROC was used to select the optimal model using the largest value.
The final values used for the model were fL = 3, usekernel = TRUE and
 adjust = 2.
          Length Class      Mode
apriori    2     table      numeric
tables    10     -none-     list
levels     2     -none-     character
call       6     -none-     call
x         10     data.frame list
usekernel  1     -none-     logical
varnames  10     -none-     character
```

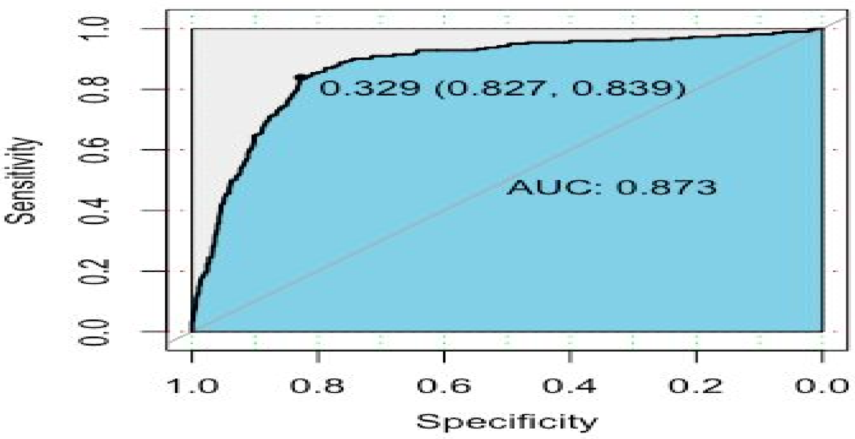
```
xNames      10     -none-     character
problemType  1     -none-     character
tuneValue    3     data.frame list
obsLevels    2     -none-     character
param        0     -none-     list

pred_class  False.  True.
    False    1353    113
    True       90    111
[1] "threshold"
[1] 0.3290708
Confusion Matrix and Statistics

          Reference
Prediction  False.  True.
   False.    1193    250
   True.       36    188

               Accuracy : 0.8284
                 95% CI : (0.8095, 0.8462)
    No Information Rate : 0.7373
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4745
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9707
            Specificity : 0.4292
         Pos Pred Value : 0.8267
         Neg Pred Value : 0.8393
             Prevalence : 0.7373
         Detection Rate : 0.7157
   Detection Prevalence : 0.8656
      Balanced Accuracy : 0.7000

       'Positive' Class :  False.
```



Naive bayes Without outliers

```
Naive Bayes

3333 samples
  10 predictor
   2 classes: 'False', 'True'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3000, 3000, 3000, 2999, 3000, 3000, ...
Resampling results across tuning parameters:

  fL  adjust  ROC        Sens       Spec
  0   0.0     0.0000000        NaN        NaN
  1   0.5     0.8309247  0.8849409  0.6154091
  3   2.0     0.8449915  0.9232861  0.5527183

Tuning parameter 'usekernel' was held constant at a value of TRUE
ROC was used to select the optimal model using the largest value.
The final values used for the model were fL = 3, usekernel = TRUE and
 adjust = 2.
          Length Class      Mode
apriori      2   table      numeric
tables      10   -none-     list
levels       2   -none-     character
call         6   -none-     call
x           10   data.frame list
usekernel    1   -none-     logical
varnames    10   -none-     character
```

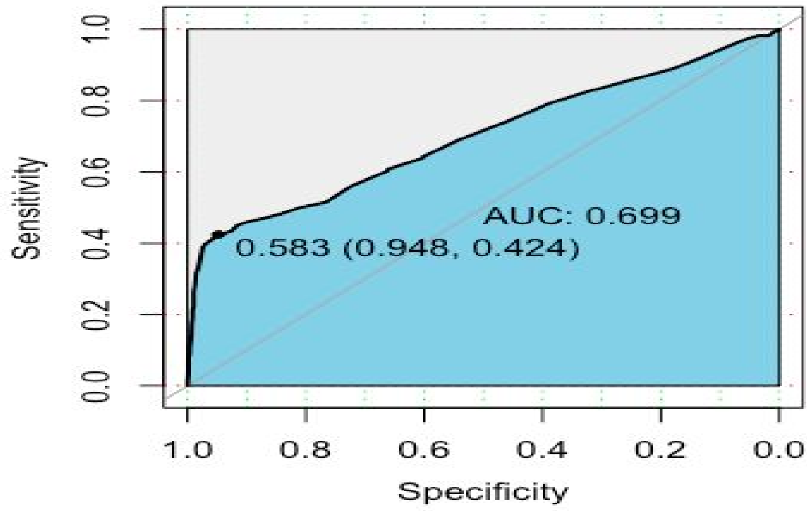
```
xNames      10     -none-     character
problemType  1     -none-     character
tuneValue    3     data.frame list
obsLevels    2     -none-     character
param        0     -none-     list

pred_class  False.  True.
    False    1353    113
    True       90    111
[1] "threshold"
[1] 0.3290708
Confusion Matrix and Statistics

          Reference
Prediction  False.  True.
    False.    1193    250
    True.       36    188

               Accuracy : 0.8284
                 95% CI : (0.8095, 0.8462)
    No Information Rate : 0.7373
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4745
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9707
            Specificity : 0.4292
         Pos Pred Value : 0.8267
         Neg Pred Value : 0.8393
             Prevalence : 0.7373
         Detection Rate : 0.7157
   Detection Prevalence : 0.8656
      Balanced Accuracy : 0.7000

       'Positive' Class :  False.
```



Knn algorithm with outliers

```
k-Nearest Neighbors

3333 samples
  10 predictor
   2 classes: 'False', 'True'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3000, 2999, 3000, 3000, 3000, 3000, ...
Resampling results across tuning parameters:

  k   ROC        Sens       Spec
  50  0.6658567  0.8280179  0.4404570
  51  0.6650417  0.8297621  0.4385937
  52  0.6657257  0.8338285  0.4354996
  53  0.6662192  0.8332471  0.4355072
  54  0.6666099  0.8326657  0.4398436
  55  0.6667708  0.8355727  0.4379917
  56  0.6678151  0.8268416  0.4379917
  57  0.6704321  0.8396357  0.4342688
  58  0.6704069  0.8349845  0.4361284
  59  0.6712957  0.8384729  0.4348785
  60  0.6705604  0.8361507  0.4355072
  61  0.6714841  0.8402137  0.4324017
  62  0.6741665  0.8396357  0.4354919
  63  0.6740760  0.8448649  0.4305345
  64  0.6740434  0.8361574  0.4311518
  65  0.6741820  0.8408019  0.4324055
  66  0.6753068  0.8384763  0.4324055
  67  0.6750814  0.8373168  0.4299172
  68  0.6746330  0.8379016  0.4274327
  69  0.6745474  0.8454530  0.4255694
  70  0.6740567  0.8437055  0.4255732
  71  0.6736705  0.8437088  0.4261943
  72  0.6739434  0.8384763  0.4311633
  73  0.6723616  0.8425393  0.4299249
  74  0.6723881  0.8378915  0.4255732
  75  0.6736135  0.8419546  0.4261943
  76  0.6726435  0.8425360  0.4255770
  77  0.6724118  0.8425360  0.4224676
  78  0.6727696  0.8454429  0.4230887
  79  0.6730288  0.8460243  0.4243348
  80  0.6736274  0.8442835  0.4212292

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 66.
           Length Class      Mode
learn        2     -none-     list
k            1     -none-     numeric
theDots      0     -none-     list
xNames      10     -none-     character
problemType  1     -none-     character
tuneValue    1     data.frame list
obsLevels    2     -none-     character
param        0     -none-     list

pred_class  False.  True.
    False    1238    117
    True      205    107
[1] "threshold"
[1] 0.4285391
Confusion Matrix and Statistics

         Reference
Prediction  False.  True.
   False.     955    488
   True.       88    136
```

```
               Accuracy : 0.6545
                 95% CI : (0.6311, 0.6773)
    No Information Rate : 0.6257
    P-Value [Acc > NIR] : 0.007871

                  Kappa : 0.1533
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9156
            Specificity : 0.2179
         Pos Pred Value : 0.6618
         Neg Pred Value : 0.6071
             Prevalence : 0.6257
         Detection Rate : 0.5729
   Detection Prevalence : 0.8656
      Balanced Accuracy : 0.5668

       'Positive' Class :  False.
```



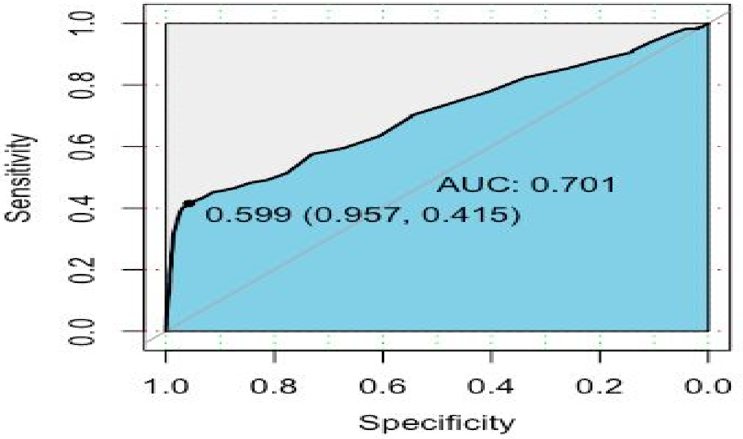
KNN-Without outliers

k-Nearest Neighbors

```
3333 samples
  10 predictor
   2 classes: 'False', 'True'

No pre-processing
Resampling: Cross-Validated (10 fold)
```

```
Summary of sample sizes: 3000, 2999, 2999, 2999, 3000, 3000, ...
Resampling results across tuning parameters:

  k   ROC        Sens       Spec
  50  0.6658628  0.8233499  0.4367456
  51  0.6658669  0.8256822  0.4355111
  52  0.6655886  0.8239414  0.4367572
  53  0.6658167  0.8239380  0.4361322
  54  0.6677144  0.8297453  0.4392301
  55  0.6687069  0.8210311  0.4398551
  56  0.6696140  0.8262535  0.4410935
  57  0.6710889  0.8262569  0.4423319
  58  0.6704100  0.8285858  0.4336401
  59  0.6705680  0.8314861  0.4348823
  60  0.6708760  0.8367119  0.4398436
  61  0.6723370  0.8355559  0.4336439
  62  0.6728698  0.8367119  0.4330151
  63  0.6749586  0.8384595  0.4361169
  64  0.6738397  0.8384628  0.4348785
  65  0.6750099  0.8396256  0.4330074
  66  0.6753893  0.8396290  0.4293076
  67  0.6761143  0.8384696  0.4348938
  68  0.6769205  0.8372967  0.4317920
  69  0.6766792  0.8390442  0.4317920
  70  0.6762539  0.8384662  0.4342842
  71  0.6758101  0.8390409  0.4286903
  72  0.6750313  0.8396256  0.4231041
  73  0.6742751  0.8384595  0.4237175
  74  0.6743310  0.8425292  0.4268269
  75  0.6741960  0.8448515  0.4212407
  76  0.6741106  0.8448548  0.4218657
  77  0.6745419  0.8460142  0.4218618
  78  0.6751425  0.8431073  0.4231079
  79  0.6749473  0.8465956  0.4218618
  80  0.6748462  0.8471804  0.4193889

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 68.
        Length Class  Mode
learn     2    -none- list
k         1    -none- numeric
theDots   0    -none- list
xNames   10    -none- character
```

```
problemType  1     -none-     character
tuneValue    1     data.frame list
obsLevels    2     -none-     character
param        0     -none-     list

pred_class  False.  True.
    False    1241    119
    True      202    105
[1] "threshold"
[1] 0.4525789
Confusion Matrix and Statistics

          Reference
Prediction  False.  True.
    False.    1054    389
    True.       95    129

               Accuracy : 0.7097
                 95% CI : (0.6872, 0.7314)
    No Information Rate : 0.6893
    P-Value [Acc > NIR] : 0.0375

                  Kappa : 0.1971
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9173
            Specificity : 0.2490
         Pos Pred Value : 0.7304
         Neg Pred Value : 0.5759
             Prevalence : 0.6893
         Detection Rate : 0.6323
   Detection Prevalence : 0.8656
      Balanced Accuracy : 0.5832

       'Positive' Class :  False.
```


AUC: 0.701
0.599 (0.957, 0.415)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

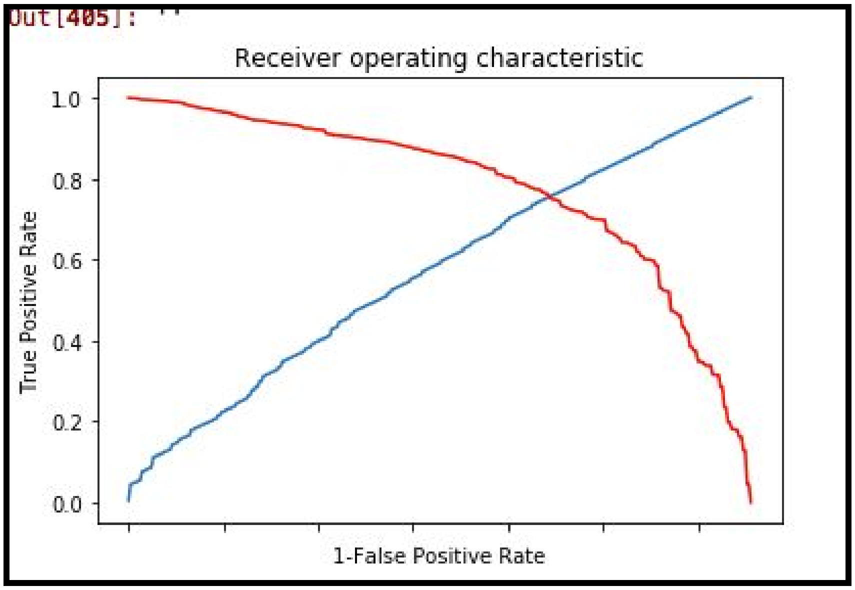
Chapter -2

Python code sample output

Logistic regression(with outliers)

```
cross_valid(logreg,test,train)
confusion matrix  of predictions(not probability predictions) and model with out using cross
validation
[[1193  250]
 [ 106  118]]
accuracy of cross validation predictions(not probability predition)
0.871025794841
mean accuracy score of  10 cross validations
0.821612091681
predicted probabilities from cross validation
          0        1
0    0.958900  0.041100
1    0.956604  0.043396
2    0.897268  0.102732
3    0.712763  0.287237
4    0.794802  0.205198
5    0.998180  0.001820
.
.
.
1657  0.821052  0.178948
1658  0.775808  0.224192
1659  0.981232  0.018768
1660  0.993518  0.006482
1661  0.965620  0.034380
1662  0.855802  0.144198
1663  0.986137  0.013863
1664  0.928676  0.071324
1665  0.979755  0.020245
1666  0.973723  0.026277

[1667 rows x 2 columns]
area under the curve
0.820605633106
```

```
threshold obtained from ROC curve
0.135393677942

confusion matrix for predictions(after applying threshold to probabilities) and test data
[[1086  357]
 [  55  169]]
accuracy of cross validation predictions( probability prediction)
0.752849430114
Out[405]: ''
```

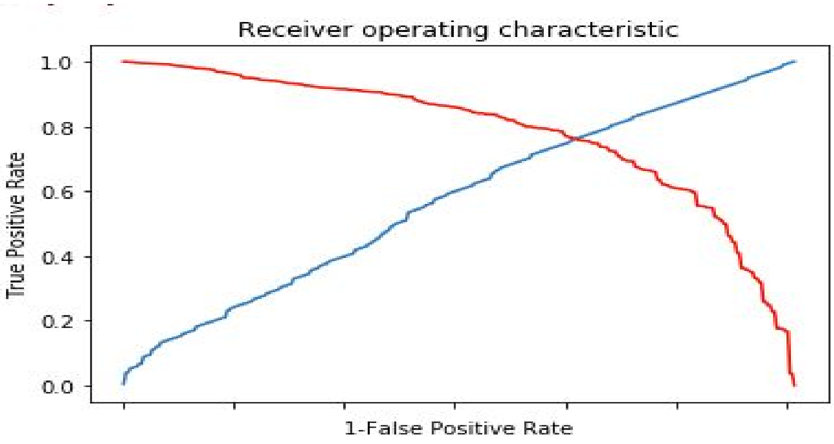


```
Out[405]: ''
```

Logistic regression-Without outliers(clean data)

```
cross_valid(logreg,test,clean_train)
confusion matrix  of predictions(not probability predictions) and model with out using cross validation
[[1189  254]
 [ 105  119]]
accuracy of cross validation predictions(not probability predition)
0.872825434913
mean accuracy score of  10 cross validations
0.825278146911
predicted probabilities from cross validation
          0         1
0    0.831928  0.168072
1    0.883816  0.116184
```

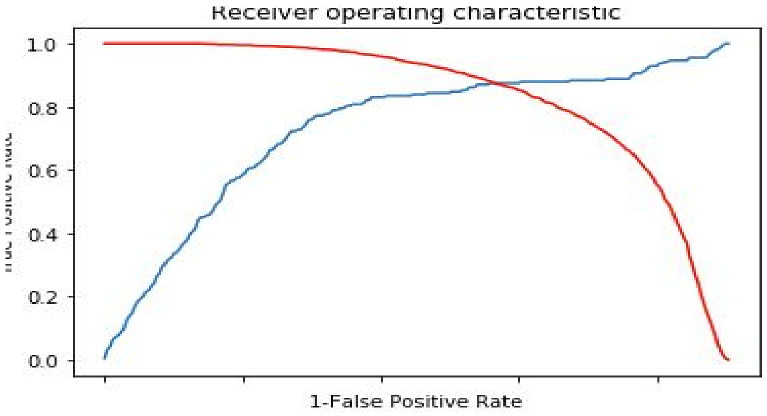
```
2     0.815944  0.184056
3     0.962401  0.03759
      ...       ...
1665  0.843911  0.156089
1666  0.947857  0.052143

[1667 rows x 2 columns]
area under the curve
0.824689387189
threshold obtained from ROC curve
0.139336951187
confusion matrix for predictions(after applying threshold to probabilities) and test data
[[1099  344]
 [  54  170]]
accuracy of cross validation predictions( probability prediction)
0.76124775045
```



Random forest with outliers

```
With outliers
cross_valid(rf,test,train)
confusion matrix  of predictions(not probability predictions) and model with out using cross validation
[[1421   22]
 [ 140   84]]
accuracy of cross validation predictions(not probability predition)
0.941211757648
mean accuracy score of  10 cross validations
0.919777601729
```

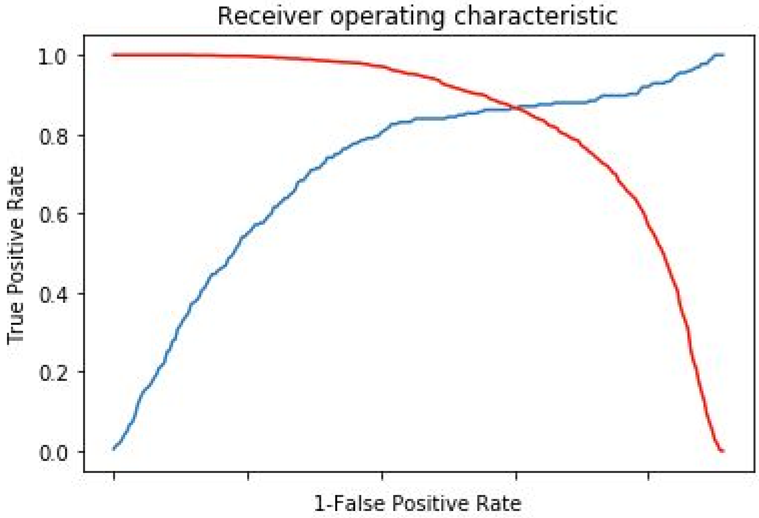
```
predicted probabilities from cross validation
         0     1
0     0.950  0.050
1     0.946  0.054
2     0.954  0.046
1665  0.304  0.696
1666  0.984  0.016

[1667 rows x 2 columns]
area under the curve
0.919313372438
threshold obtained from ROC curve
0.176
confusion matrix for predictions(after applying threshold to probabilities) and test data
[[1266  177]
 [  28  196]]
accuracy of cross validation predictions( probability prediction)
0.877024595081
```



Random forestWith out outliers (cleaned data)

```
cross_valid(rf,test,clean_train)
confusion matrix  of predictions(not probability predictions) and model with out using cross validation
[[1402   41]
 [ 136   88]]
accuracy of cross validation predictions(not probability predition)
0.944211157768
mean accuracy score of  10 cross validations
0.920495935744
predicted probabilities from cross validation
```

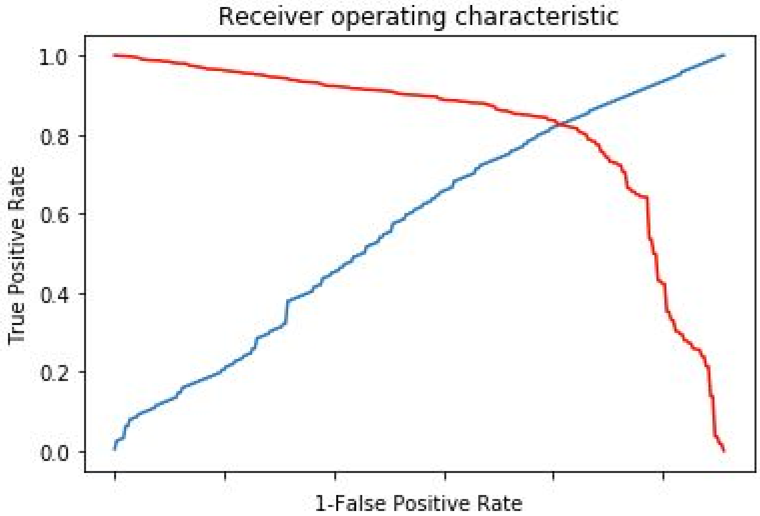
```
        0      1
0     0.938  0.062
1     0.942  0.058
1666  0.976  0.024

[1667 rows x 2 columns]
area under the curve
0.920309560935
threshold obtained from ROC curve
0.168
confusion matrix for predictions(after applying threshold to probabilities) and test data
[[1251  192]
 [  30  194]]
accuracy of cross validation predictions( probability prediction)
0.866826634673
```



Naive bayes algorithm-With outliers

```
cross_valid(gnb,test,train)
confusion matrix  of predictions(not probability predictions) and model with out using cross validation
[[1340  103]
 [ 151   73]]
accuracy of cross validation predictions(not probability predition)
0.862627474505
mean accuracy score of  10 cross validations
0.860475690185
```

```
predicted probabilities from cross validation
          0        1
0     0.958844  0.041156
1     0.949762  0.050238
2     0.939692  0.060308


1666  0.997729  0.002271

[1667 rows x 2 columns]
area under the curve
0.860614048114
threshold obtained from ROC curve
0.122283585447
confusion matrix for predictions(after applying threshold to probabilities) and test data
[[1194  249]
 [  40  184]]
accuracy of cross validation predictions( probability prediction)
0.826634673065
```

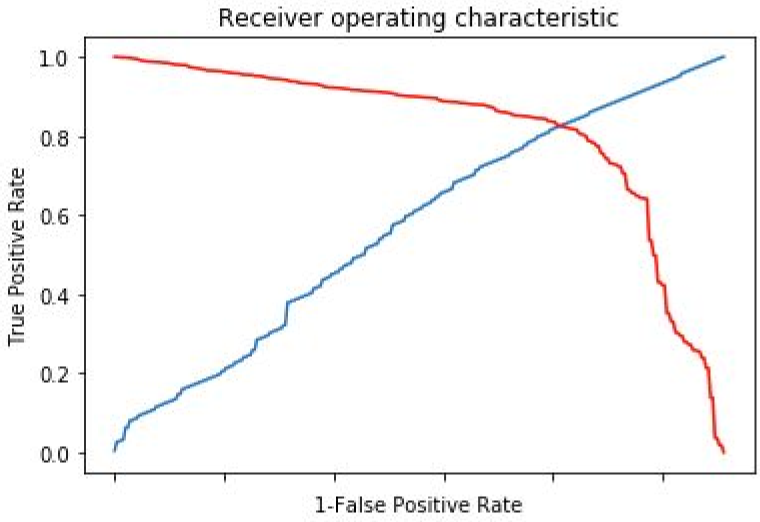


… …

Naive bayes without outliers[cleaned]

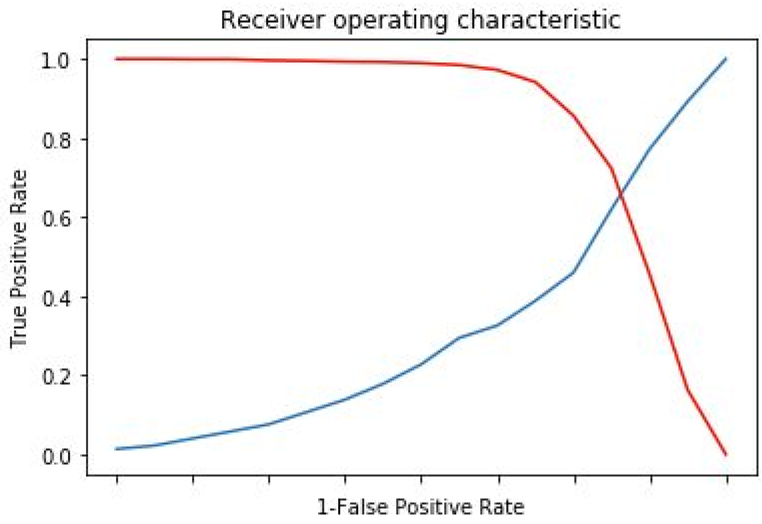
```
cross_valid(gnb,test,clean_train)
confusion matrix  of predictions(not probability predictions) and model with out using cross validation
[[1277  166]
 [ 130   94]]
accuracy of cross validation predictions(not probability predition)
```

```
0.862627474505
mean accuracy score of  10 cross validations
0.860475690185
predicted probabilities from cross validation
           0        1
0     0.958844  0.041156
1     0.949762  0.050238

       ...      ...

1666  0.997729  0.002271

[1667 rows x 2 columns]
area under the curve
0.860614048114
threshold obtained from ROC curve
0.122283585447
confusion matrix for predictions(after applying threshold to probabilities) and test data
[[1194  249]
 [  40  184]]
accuracy of cross validation predictions( probability prediction)
0.826634673065
```



KNN model-With outliers

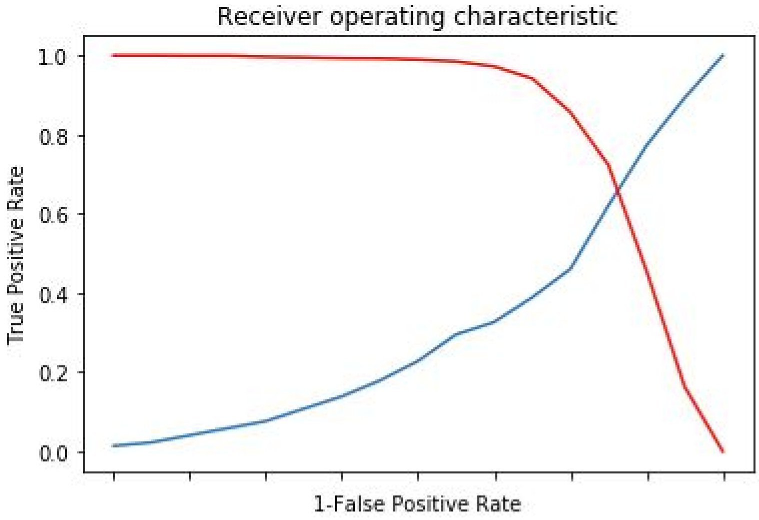
```
cross_valid(kNN,test,train)
confusion matrix  of predictions(not probability predictions) and model with out using cross validation
[[1435    8]
 [ 196   28]]
accuracy of cross validation predictions(not probability predition)
0.875824835033
mean accuracy score of  10 cross validations
0.710682031144
predicted probabilities from cross validation
        0     1
0     1.00  0.00
1665  0.85  0.15
1666  0.80  0.20

[1667 rows x 2 columns]
area under the curve
0.709600225225
threshold obtained from ROC curve
0.15
confusion matrix for predictions(after applying threshold to probabilities) and test data
[[1236  207]
 [ 121  103]]
accuracy of cross validation predictions( probability prediction)
0.80323935213
```



KNN -without outliers

```
cross_valid(kNN,test,clean_train)
confusion matrix  of predictions(not probability predictions) and model with out using cross validation
[[958 485]
 [ 94 130]]
accuracy of cross validation predictions(not probability predition)
0.875824835033
mean accuracy score of  10 cross validations
0.710682031144
predicted probabilities from cross validation
        0     1
0     1.00  0.00
1     0.90  0.10
1665  0.85  0.15
1666  0.80  0.20

[1667 rows x 2 columns]
area under the curve
0.709600225225
threshold obtained from ROC curve
0.15
confusion matrix for predictions(after applying threshold to probabilities) and test data
[[1236  207]
 [ 121  103]]
accuracy of cross validation predictions( probability prediction)
```

```
0.80323935213
```


Receiver operating characteristic
True Positive Rate
1-False Positive Rate