FIT3152 Data Analytics

Assignment 2

"Gain Familiarity with Classification Models Using R"

Loi Chai Lam - 28136179

Data Exploration

At the beginning, exploration of the Kaggle competition data given had been done. The proportion of rainy days to fine days is 2085 : 7771, which shows that fine days are significantly larger than rainy days. Two summary tables[Table1][Table2] for the desciptions of the predictor (independent) variables was shown below:

Table 1 : The mean and standard deviation of the data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Stand.Dev.
Day	1.0	8.0	16.0	15.659123	23.0	31.0	8.829681
Month	1.0	4.0	6.0	6.445332	9.0	12.0	3.419047
Year	2008.0	2011.0	2013.0	2012.780714	2015.0	2017.0	2.547269
Location	2.0	7.0	14.0	15.692800	26.0	37.0	10.853249
MinTemp	-5.2	7.9	13.0	13.081973	18.0	30.3	6.692789
MaxTemp	7.2	18.6	23.3	24.130829	30.0	46.7	7.128947
Rainfall	0.0	0.0	0.0	2.188469	0.6	162.2	7.963563
Evaporation	0.0	2.8	5.0	5.491346	7.2	48.8	3.763428
Sunshine	0.0	5.6	8.8	7.858498	10.7	14.0	3.595712
WindGustSpeed	9.0	31.0	37.0	39.692466	46.0	120.0	13.092719
WindSpeed9am	0.0	9.0	13.0	14.407148	20.0	65.0	8.220752
WindSpeed3pm	0.0	13.0	17.0	18.135986	22.0	59.0	7.887264
Humidity9am	3.0	54.0	67.0	65.161552	79.0	100.0	19.530566
Humidity3pm	1.0	31.0	50.0	48.301887	63.0	100.0	21.395781
Pressure9am	980.5	1012.2	1016.8	1016.813873	1021.8	1039.3	7.285748
Pressure3pm	977.1	1009.4	1014.3	1014.324537	1019.3	1036.3	7.236540
Cloud9am	0.0	1.0	5.0	4.195879	7.0	8.0	2.837525
Cloud3pm	0.0	2.0	5.0	4.251138	7.0	8.0	2.664384
Temp9am	-0.2	12.9	17.6	17.997857	22.9	38.2	6.731694
Temp3pm	4.8	17.3	22.0	22.745827	28.3	46.1	7.054347

Table 2 : The number of NAs in the data.

```
      Day             Month             Year          Location       MinTemp          MaxTemp         Rainfall        Evaporation    
 Min.   : 1.00   Min.   : 1.000   Min.   :2008   Min.   : 2.00   Min.   :-5.20   Min.   : 7.20   Min.   :  0.000   Min.   : 0.000  
 1st Qu.: 8.00   1st Qu.: 4.000   1st Qu.:2011   1st Qu.: 7.00   1st Qu.: 7.90   1st Qu.:18.60   1st Qu.:  0.000   1st Qu.: 2.800  
 Median :16.00   Median : 6.000   Median :2013   Median :14.00   Median :13.00   Median :23.30   Median :  0.000   Median : 5.000  
 Mean   :15.66   Mean   : 6.445   Mean   :2013   Mean   :15.69   Mean   :13.08   Mean   :24.13   Mean   :  2.188   Mean   : 5.491  
 3rd Qu.:23.00   3rd Qu.: 9.000   3rd Qu.:2015   3rd Qu.:26.00   3rd Qu.:18.00   3rd Qu.:30.00   3rd Qu.:  0.600   3rd Qu.: 7.200  
 Max.   :31.00   Max.   :12.000   Max.   :2017   Max.   :37.00   Max.   :30.30   Max.   :46.70   Max.   :162.200   Max.   :48.800  
 NA's   :146     NA's   :113      NA's   :86                     NA's   :126     NA's   :110     NA's   :140       NA's   :2535    
    Sunshine      WindGustDir   WindGustSpeed     WindDir9am     WindDir3pm    WindSpeed9am    WindSpeed3pm    Humidity9am    
 Min.   : 0.000   E      : 770   Min.   :  9.00   SE     : 799   SE     : 892   Min.   : 0.00   Min.   : 0.00   Min.   :  3.00  
 1st Qu.: 5.600   SSE    : 711   1st Qu.: 31.00   NNW    : 775   ESE    : 797   1st Qu.: 9.00   1st Qu.:13.00   1st Qu.: 54.00  
 Median : 8.800   SE     : 694   Median : 37.00   E      : 751   E      : 675   Median :13.00   Median :17.00   Median : 67.00  
 Mean   : 7.858   ESE    : 625   Mean   : 39.69   ESE    : 663   NNW    : 675   Mean   :14.41   Mean   :18.14   Mean   : 65.16  
 3rd Qu.:10.700   NW     : 613   3rd Qu.: 46.00   SSE    : 638   WNW    : 646   3rd Qu.:20.00   3rd Qu.:22.00   3rd Qu.: 79.00  
 Max.   :14.000   (Other):5363   Max.   :120.00   (Other):5922   (Other):5911   Max.   :65.00   Max.   :59.00   Max.   :100.00  
 NA's   :3234     NA's   :1224   NA's   :1240     NA's   : 452   NA's   : 404   NA's   :151     NA's   :374     NA's   :127     
  Humidity3pm     Pressure9am     Pressure3pm      Cloud9am        Cloud3pm        Temp9am         Temp3pm       RainToday   
 Min.   :  1.0   Min.   : 980.5   Min.   : 977.1   Min.   :0.000   Min.   :0.000   Min.   :-0.2   Min.   : 4.80   No  :7771  
 1st Qu.: 31.0   1st Qu.:1012.2   1st Qu.:1009.4   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:12.9   1st Qu.:17.30   Yes :2085  
 Median : 50.0   Median :1016.8   Median :1014.3   Median :5.000   Median :5.000   Median :17.6   Median :22.00   NA's: 144  
 Mean   : 48.3   Mean   :1016.8   Mean   :1014.3   Mean   :4.196   Mean   :4.251   Mean   :18.0   Mean   :22.75              
 3rd Qu.: 63.0   3rd Qu.:1021.8   3rd Qu.:1019.3   3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:22.9   3rd Qu.:28.30              
 Max.   :100.0   Max.   :1039.3   Max.   :1036.3   Max.   :8.000   Max.   :8.000   Max.   :38.2   Max.   :46.10              
 NA's   :354     NA's   :1105     NA's   :1091     NA's   :2526    NA's   :2753    NA's   :109    NA's   :366                
RainTomorrow
No  :7900
Yes :1999
NA's: 101
```

From the summary table [Table2] above, there are many missing values (NA) in the data. The missing data in the data set can reduce the power of a model or can lead to a biased model because the behavior and relationship with other variables have not been analyzed correctly. It can lead to wrong prediction or classification.

Pre-processing

From the data exploration above, Day, Month and Year variables are decided to be removed in the data set. This is because the model is predicting whether it will rain tomorrow from all the data given, without considering seasonality, or the date. Hence, with that pre-processing, the model can focus more on the predictor variables which might be more important to the model. Also, the default data structure of "Location" variable had been changed from integer to factor, as Location is not a continuous number and it is just a limited of number. Lastly, all the missing values in the data had been removed.

Classification Model

After creating the classification models, each of the test cases is classified as "will rain tomorrow" or "will not rain tomorrow" using the test data. The confusion matrixes, the accuracy and the AUC for each classification models are shown below.

Decision Tree

```
> print(WAUS.dtree.confusion$table)
                   predicted_RainTomorrrow
actual_RainTomorrow  No Yes
                No  957  58
                Yes 142 103

> cat ("Accuracy is :",WAUS.dtree.confusion$overall["Accuracy"])
Accuracy is : 0.8412698

> cat("AUC is ",as.numeric(AUC.dtree@y.values))
AUC is  0.8309601
```

Naïve Bayes

```
> print(WAUS.nbayes.confusion$table)
                    predicted_RainTomorrow
actual_RainTomorrow  No Yes
                No  854 161
                Yes  85 160
> cat ("Accuracy is :",WAUS.nbayes.confusion$overall["Accuracy"])
Accuracy is : 0.8047619

> cat("AUC is ",as.numeric(AUC.nbayes@y.values))
AUC is  0.8514185
```

Bagging

```
> print(WAUS.bagging.confusion$table)
                    predicted_RainTomorrow
actual_RainTomorrow  No Yes
                No  959  56
                Yes 123 122

> cat ("Accuracy is :",WAUS.bagging.confusion$overall["Accuracy"])
Accuracy is : 0.8579365

> cat("AUC is ",as.numeric(AUC.bagging@y.values))
AUC is  0.8056017
```

Boosting

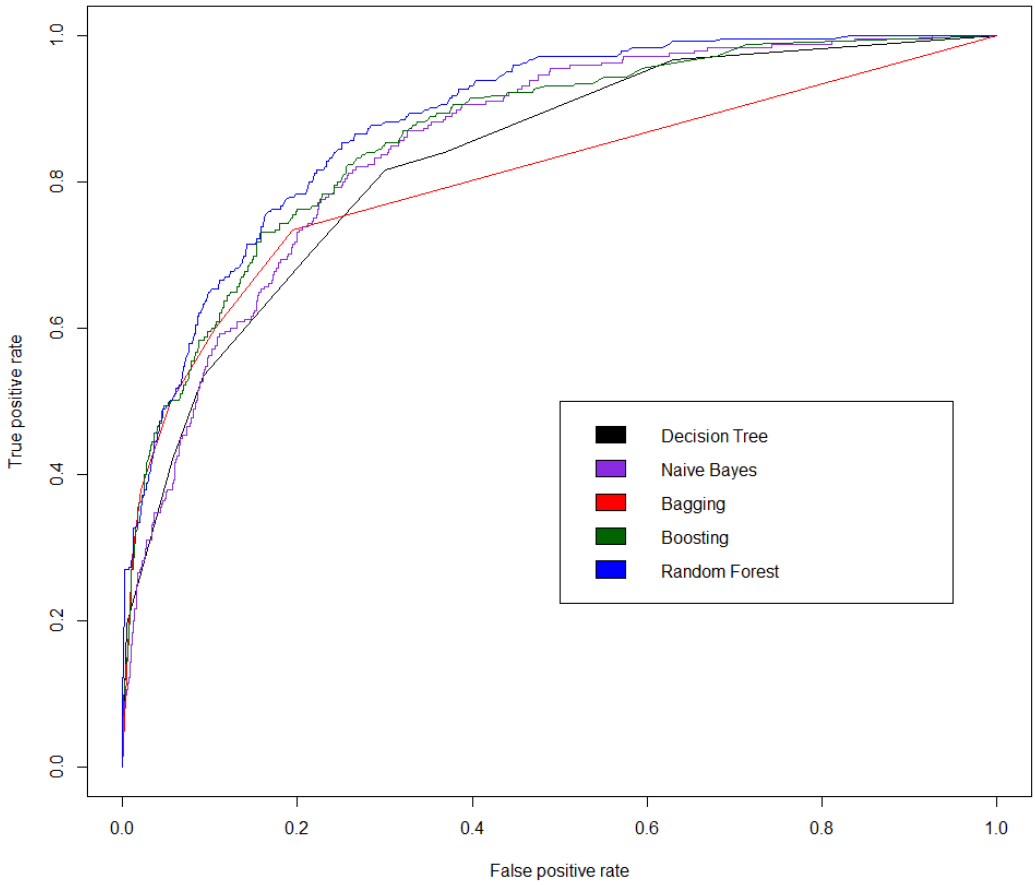
```
> print(WAUS.boosting.confusion$table)
                    predicted_RainTomorrow
actual_RainTomorrow  No Yes
                No  933  82
                Yes 109 136
> cat ("Accuracy is :",WAUS.boosting.confusion$overall["Accuracy"])
Accuracy is : 0.8484127

> cat("AUC is ",as.numeric(AUC.boosting@y.values))
AUC is  0.8637459
```

Random Forest

```
> print(WAUS.rforests.confusion$table)
      predicted_RainTomorrrow
actual  No Yes
   No  935  80
   Yes 103 142
> cat ("Accuracy is :",WAUS.rforests.confusion$overall["Accuracy"])
Accuracy is : 0.8547619
```

```
> cat("AUC is ",as.numeric(AUC.rforests@y.values))
AUC is  0.8845823
```

The ROC curve for each classifier is shown as below.



Graph1: ROC curve for each classifier

Classification Model Analysis

A table [Table 3] is created to compare all the results of all classifiers.

Table 3: The results of all classifiers

```
> show(showTable)
     [,1]          [,2]               [,3]
[1,] Model         Accuracy           AUC
[2,] Decision Tree 0.841269841269841  0.830960088468885
[3,] Naive Bayes   0.804761904761905  0.851418518146168
[4,] Bagging       0.857936507936508  0.805601688951443
[5,] Boosting      0.848412698412698  0.863745853021011
[6,] Random Forest 0.854761904761905  0.884582286116417
```

From the table [Table3] above, we can conclude that Random Forest model is the single best classifier for all. The accuracy of Random Forest is 0.855, which is slightly lower than Bagging. However, its AUC value is 0.885, which is the highest among the others.

Variable Analysis

The most important variables in predicting whether or not it will rain tomorrow for each of the classification models are shown below.

Decision Tree

```
#Decision Tree Attribute Importance
> summary(WAUS.dtree)

Classification tree:
tree(formula = RainTomorrow ~ ., data = WAUS.train)
Variables actually used in tree construction:
[1] "Humidity3pm"   "Sunshine"      "Pressure3pm"   "MinTemp"       "Location"      "WindGustSpeed"
Number of terminal nodes:  8
Residual mean deviance:  0.7074 = 2072 / 2929
Misclassification error rate: 0.1593 = 468 / 2937
>
```

For Decision Tree, the most important variables are Humidity3pm, Sunshine, Pressure3pm, MinTemp, Location and WindGustSpeed.

Naïve Bayes

Naive Bayes does not provide attribute importance information. Hence, we are not able to provide any sorts of systematic analytic on the attribute importance.

Bagging

```
> print(sort(WAUS.bagging$importance,decreasing=TRUE))
  Humidity3pm      Sunshine   WindGustDir   Pressure3pm    WindDir3pm WindGustSpeed   Pressure9am    WindDir9am      Location
   44.0832049    17.1727640     6.7763328     5.8866774     5.6470004     5.5517984     3.9798735     3.0727928     2.9927279
      MaxTemp  Humidity9am       Temp3pm      Cloud3pm       Temp9am      Cloud9am  WindSpeed9am   Evaporation       MinTemp
    1.4468836     0.8514838     0.8323811     0.7977535     0.3490887     0.2986318     0.2606053     0.0000000     0.0000000
     Rainfall     RainToday  WindSpeed3pm
    0.0000000     0.0000000     0.0000000
```

For Bagging, the six most important variables are Humidity3pm, Sunshine, WindGustDir, Pressure3pm, WindDir3pm and WindGustSpeed. From the table above, Evaporation, MinTemp, Rainfall, RainToday and WindSpeed3pm could be omitted from the data because they have zero effect on the performance.

Boosting

```
> print(sort(WAUS.boosting$importance,decreasing=TRUE))
  Humidity3pm    WindDir9am    WindDir3pm      Sunshine   WindGustDir   Pressure3pm      Cloud3pm WindGustSpeed   Pressure9am
   27.8437285    13.2479669    10.1912220     9.7864983     8.8101716     5.5259940     3.2636549     3.2041550     2.7919799
      MinTemp      Location  Humidity9am   Evaporation       MaxTemp      Rainfall       Temp9am       Temp3pm  WindSpeed9am
    2.4588800     2.3035510     1.8535772     1.8112780     1.7840487     1.3303121     1.1902155     0.8574497     0.8024823
WindSpeed3pm      Cloud9am     RainToday
   0.5910007     0.3518339     0.0000000
```

For Boosting, the six most important variables are Humidity3pm, WindDir9am, WindDir3pm, Sunshine, WindGustDir and Pressure3pm. From the table above, RainToday could be omitted from the data because it has zero effect on the performance.

Random Forest

```
> print(WAUS.rforests$importance)
              MeanDecreaseGini
Location              21.78486
MinTemp               37.30499
MaxTemp               31.01076
Rainfall              34.82992
Evaporation           28.71874
Sunshine              93.36894
WindGustDir           70.57745
WindGustSpeed         38.10049
WindDir9am            72.31138
WindDir3pm            68.56122
WindSpeed9am          23.22895
WindSpeed3pm          23.36303
Humidity9am           38.90692
Humidity3pm          135.22503
Pressure9am           52.27931
Pressure3pm           51.75078
Cloud9am              28.97285
Cloud3pm              52.60990
Temp9am               33.02788
Temp3pm               31.53555
RainToday             16.00864
```

For Random Forest, the six most important variables are Humidity3pm, Sunshine, WinDir9am, WindGustDir, WindDir3pm and Cloud3pm. The variables that could be omitted from the data with very little effect on performance is RainToday.

From the analysis above, RainToday is the variable which less effect on the performance for most of the classification models. Meanwhile, Humidity3pm, Sunshine and WindGustDir are the most important variables for most of the classification models.

Reconstruct Classification Model

Decision Tree

For the Decision Tree, k-fold cross validation and pruning are used to determine if the original tree is overfit or underfit. The information of the k-fold cross validation is shown below.

```
> WAUS.dtree2
$size
[1] 8 4 1

$dev
[1] 492 492 625

$k
[1]      -Inf   0.00000 52.33333

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```

The original initial decision tree has 8 terminal nodes, Accuracy of 0.8413 and AUC of 0.8310.
From the information given, the recommended value for decision tree is 4 nodes. However, the deviation for 8 nodes and 4 nodes are same. After experimenting with the prune size of 4, the Accuracy remained the same as 0.8413, but the AUC values of the pruned decision tree dropped to 0.7575. The information for the pruned decision tree is shown below. Hence, the original decision tree is the best decision tree.

```
> WAUS.dtree.prune.confusion
Confusion Matrix and Statistics

      predicted
actual  No Yes
   No  957  58
   Yes 142 103

               Accuracy : 0.8413
                 95% CI : (0.8199, 0.861)
    No Information Rate : 0.8722
    P-Value [Acc > NIR] : 0.9994

                  Kappa : 0.4176

 Mcnemar's Test P-Value : 4.385e-09

            Sensitivity : 0.8708
            Specificity : 0.6398
         Pos Pred Value : 0.9429
         Neg Pred Value : 0.4204
             Prevalence : 0.8722
         Detection Rate : 0.7595
   Detection Prevalence : 0.8056
      Balanced Accuracy : 0.7553

       'Positive' Class : No

> print(as.numeric(WAUS.dtree.prune.auc@y.values))
[1] 0.7574786
```

Bagging (show graph or table of accuracy, AUC, ROC)

The original Bagging model has an Accuracy 0.8579 of and AUC of 0.8056.

To improve the performance of bagging, the value of mfinal had changed to 15, as mfinal 15 gives the best performance after experimenting with other possible values. The improved Bagging model has an Accuracy of 0.8706 and AUC of 0.8272. The model has improved by 0.0127 accuracy and 0.0216 AUC.

The information is shown below.

```
> print(WAUS.bagging.confusion2)
Confusion Matrix and Statistics

      predicted
actual  No Yes
   No  976  39
   Yes 124 121

               Accuracy : 0.8706
                 95% CI : (0.8508, 0.8887)
    No Information Rate : 0.873
    P-Value [Acc > NIR] : 0.6201

                  Kappa : 0.5245

 Mcnemar's Test P-Value : 4.724e-11

            Sensitivity : 0.8873
            Specificity : 0.7562
         Pos Pred Value : 0.9616
         Neg Pred Value : 0.4939
             Prevalence : 0.8730
         Detection Rate : 0.7746
   Detection Prevalence : 0.8056
      Balanced Accuracy : 0.8218

       'Positive' Class : No

> print(as.numeric(AUC.bagging2@y.values))
[1] 0.8272323
```

Hence, increasing the number mfinal to 15 improved the accuracy and AUC of the Bagging model.

Random Forest

The original Random Forest model has an Accuracy 0.8548 of and AUC of 0.8846.
To improve the current Random forest model, the ntree was increased to 2000 and mtry was increased to 5. Those two values improved the model to Accuracy of 0.8595 and AUC of 0.8862. The model has improved by 0.0047 accuracy and 0.0016 AUC.

The information below shown the improved Random Forest.

```
> print(WAUS.rforests.confusion2)
Confusion Matrix and Statistics

      predicted
actual  No Yes
   No  938  77
   Yes 100 145

               Accuracy : 0.8595
                 95% CI : (0.8391, 0.8783)
    No Information Rate : 0.8238
    P-Value [Acc > NIR] : 0.0003656

                  Kappa : 0.535

 Mcnemar's Test P-Value : 0.0982045

            Sensitivity : 0.9037
            Specificity : 0.6532
         Pos Pred Value : 0.9241
         Neg Pred Value : 0.5918
             Prevalence : 0.8238
         Detection Rate : 0.7444
   Detection Prevalence : 0.8056
      Balanced Accuracy : 0.7784

       'Positive' Class : No

> print(as.numeric(AUC.rforests2@y.values))
[1] 0.8862592
```

Artificial Neural Network (ANN)

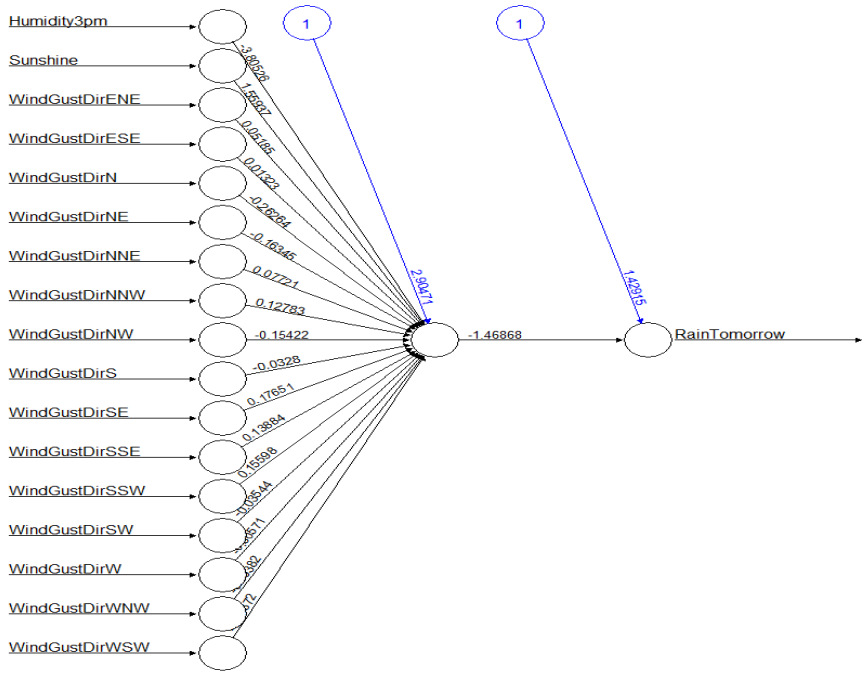
To construct an ANN model, some pre-processing of data should be completed beforehand. The categorical data should be converted to binary columns as indicator variables. The inputs can only be numerical. There should not have any missing values in the data. Also, the attributes should have been normalized for ANN to avoid the attribute with larger values to unduly affect the model more significantly than the others.

Hence, RainToday, RainTomorrow variables are converted to numeric 1 and 0 to indicate Yes or No. After that, WindGustDir, WindDir9am and WindDir3pm variables are converted to categorical matrixes. All the missing values in the data set are removed. Last, the data set are normalized.

Humidity3pm, Sunshine and WindGustDir variables are used to construct the ANN model. The reason behind this is that these three variables are the variables which are more likely to affect the prediction from the previous observation. Moreover, these three variables are used because they are able to produce the highest accuracy in the model.

The accuracy of the ANN is 0.8563, which is just slightly lower than bagging (0.8579), and slightly higher than Random Forest (0.8548). In overall, the performance of ANN is good enough as it hits the high accuracy, however, the execution time of ANN is slower than the other models, especially when the hidden layer is increased. Hence, it is not the fastest solution, but it is stable.

The constructed ANN is shown below.



Graph2: ANN model

The confusion matrix of the ANN model is shown below.

```
> print(WAUS.ann.confusion)
Confusion Matrix and Statistics

      predicted
actual   0   1
     0 982  33
     1 148  97

               Accuracy : 0.8563
                 95% CI : (0.8358, 0.8753)
    No Information Rate : 0.8968
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4421

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

            Sensitivity : 0.8690
            Specificity : 0.7462
         Pos Pred Value : 0.9675
         Neg Pred Value : 0.3959
             Prevalence : 0.8968
         Detection Rate : 0.7794
   Detection Prevalence : 0.8056
      Balanced Accuracy : 0.8076

       'Positive' Class : 0
```