MSCI 562 – Intelligent Data Analysis and Visualisation

Individual Coursework - 2

Submitted by,

36342582

Table of Contents

EXECUTIVE SUMMARY:

The objective of the analysis was to predict customer satisfaction using four classification models: Logistic Regression, KNN, Decision Tree, and Random Forest. Among these models, Random Forest demonstrated the highest level of generalization performance with an AUC value of 0.991, outperforming the other models. The influential predictor variables for customer satisfaction included Onlineboarding, Inflightwifiservice, Class, TypeofTravel, Inflightentertainment, and Customer Type.

The ROC curve provided insights into the trade-off between accurately identifying satisfied and dissatisfied customers. Based on the findings, the Random Forest model was identified as the most suitable choice for accurately predicting customer satisfaction and facilitating decision-making in the airline industry. Its strong and reliable performance makes it highly recommended for deployment.

INTRODUCTION:

This report focuses primarily on modelling and predicting customer satisfaction in the UK-based airline industry. The exploratory data analysis, revealed a slightly higher percentage of negative/neutral responses, indicating room for improvement in service quality. Further analysis using techniques such as IV, WOE, and MDS allowed us to identify key attributes, including Inflightwifiservice, OnlineBoarding etc., which significantly impact customer satisfaction.

One of the primary challenges addressed in this report is the comparison of different classification methods and the selection of the most appropriate models based on the marketing manager's objective of correctly identify 95% of dissatisfied customers while minimising the misclassification rate of satisfied customers as dissatisfied (<10%).Additionally, we visually explore the trade-off between accurately identifying satisfaction and dissatisfaction .Determining the variable importance is another important aspect considered in this report. Finally, the generalisation performance of the models is discussed.

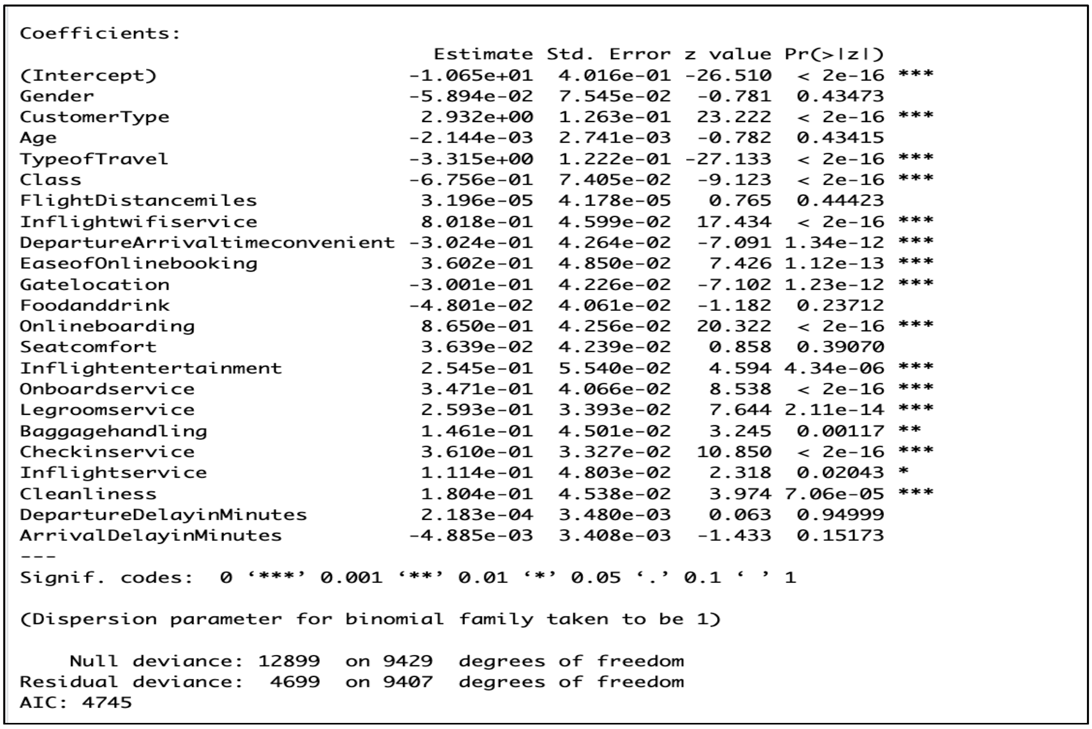
The first section of report gives an overview of various classification models and comparison of their results. Based on airline's objectives the most suitable model is selected and recommended. Then the trade-off between correct identification of clients using the decision boundary concepts is examined visually. Furthermore, the variable importance based on the prediction of satisfaction is identified and discussed. At the end, the generalisation performance of the models is discussed, utilising ROC curves for performance evaluation. Then report concludes by summarising the important findings of the given dataset and recommendation for the airline company.

CLASSIFICATION MODELS AND ANALYSIS:

To predict customer satisfaction, four different classification methods were utilised on a dataset that includes multiple independent variables, with customer satisfaction as the exogenous variable. The following paragraphs provide a comprehensive explanation of each classification model used in the analysis, outlining their suitability for the prediction task at hand.

a) Logistic Regression:

The logit model, which assumes a Bernoulli distribution for the dependent variable, is well-suited for predicting customer satisfaction as it represents binary outcome. In our analysis, we applied this model to the dataset using a threshold of 0.5 to establish a balanced decision boundary. Figure 1 below displays a summary of the model, highlighting its superior performance with the lowest AIC value of 4745 compared to other logistic regression models considered.



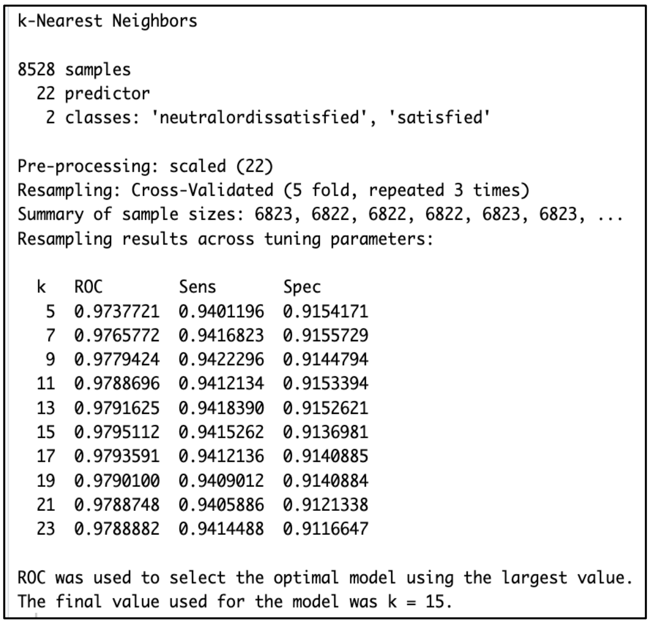
```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                    -1.065e+01  4.016e-01 -26.510  < 2e-16 ***
Gender                         -5.894e-02  7.545e-02  -0.781  0.43473    
CustomerType                    2.932e+00  1.263e-01  23.222  < 2e-16 ***
Age                            -2.144e-03  2.741e-03  -0.782  0.43415    
TypeofTravel                   -3.315e+00  1.222e-01 -27.133  < 2e-16 ***
Class                          -6.756e-01  7.405e-02  -9.123  < 2e-16 ***
FlightDistancemiles             3.196e-05  4.178e-05   0.765  0.44423    
Inflightwifiservice             8.018e-01  4.599e-02  17.434  < 2e-16 ***
DepartureArrivaltimeconvenient -3.024e-01  4.264e-02  -7.091 1.34e-12 ***
EaseofOnlinebooking             3.602e-01  4.850e-02   7.426 1.12e-13 ***
Gatelocation                   -3.001e-01  4.226e-02  -7.102 1.23e-12 ***
Foodanddrink                   -4.801e-02  4.061e-02  -1.182  0.23712    
Onlineboarding                  8.650e-01  4.256e-02  20.322  < 2e-16 ***
Seatcomfort                     3.639e-02  4.239e-02   0.858  0.39070    
Inflightentertainment           2.545e-01  5.540e-02   4.594 4.34e-06 ***
Onboardservice                  3.471e-01  4.066e-02   8.538  < 2e-16 ***
Legroomservice                  2.593e-01  3.393e-02   7.644 2.11e-14 ***
Baggagehandling                 1.461e-01  4.501e-02   3.245  0.00117 ** 
Checkinservice                  3.610e-01  3.327e-02  10.850  < 2e-16 ***
Inflightservice                 1.114e-01  4.803e-02   2.318  0.02043 *  
Cleanliness                     1.804e-01  4.538e-02   3.974 7.06e-05 ***
DepartureDelayinMinutes         2.183e-04  3.480e-03   0.063  0.94999    
ArrivalDelayinMinutes          -4.885e-03  3.408e-03  -1.433  0.15173    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12899  on 9429  degrees of freedom
Residual deviance:  4699  on 9407  degrees of freedom
AIC: 4745
```

Figure 1: Summary of Logistic Regression Model

Among the predictor variables, Customer Type, Type of Travel, Class, Inflightwifiservice, Online Boarding, and Inflightentertainment show p-values below 0.05, suggesting their strong association with customer satisfaction. However, Gender, Age, and FoodandDrink do not demonstrate a significant impact on customer satisfaction in this model.

b) K - Nearest Neighbours:

The selection of the optimal value of k is determined through cross-validation based on the random sampling with 5 folds for 10 values of k. Performance measures such as AUC (ROC), sensitivity, and specificity are used to assess the model's performance for each value of k. Since the dataset is imbalanced, SMOTE approach is utilised to address this issue. Figure 2 displays the optimal model selected based on the highest ROC value. In this case, the model with a tuning parameter (k) of 15 exhibits the best overall performance, as it achieves the maximum ROC value.



```
k-Nearest Neighbors

8528 samples
  22 predictor
   2 classes: 'neutralordissatisfied', 'satisfied'

Pre-processing: scaled (22)
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 6823, 6822, 6822, 6822, 6823, 6823, ...
Resampling results across tuning parameters:

  k   ROC        Sens       Spec
   5  0.9737721  0.9401196  0.9154171
   7  0.9765772  0.9416823  0.9155729
   9  0.9779424  0.9422296  0.9144794
  11  0.9788696  0.9412134  0.9153394
  13  0.9791625  0.9418390  0.9152621
  15  0.9795112  0.9415262  0.9136981
  17  0.9793591  0.9412136  0.9140885
  19  0.9790100  0.9409012  0.9140884
  21  0.9788748  0.9405886  0.9121338
  23  0.9788882  0.9414488  0.9116647

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 15.
```

Figure 2: Output of KNN Selection

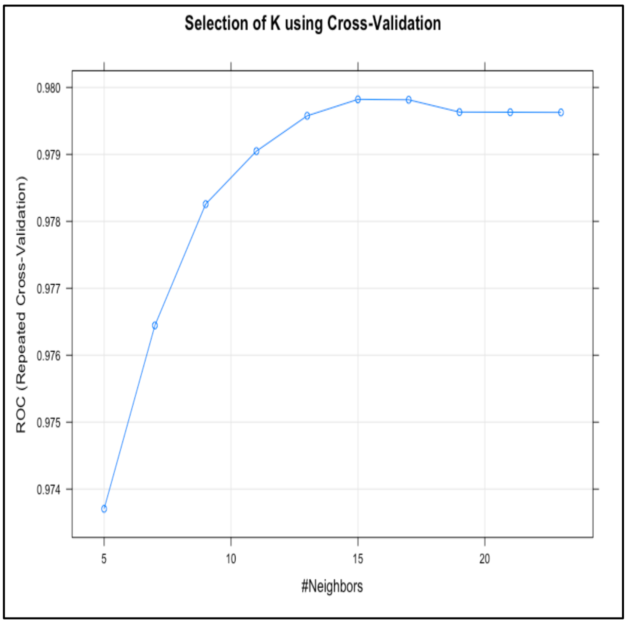


Figure 3: Plot of KNN Selection

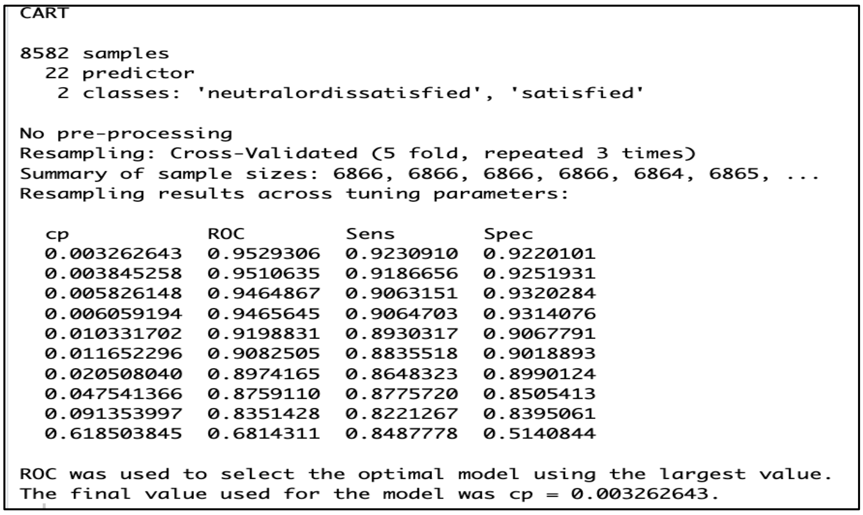
Figure 3 provides further insight into the ROC values across different values of k during cross-validation, reaffirming the selection of k=15 as the optimal choice. Finally, with the optimal k value determined, the model can be applied to the test data to predict customer satisfaction.

c) CART – Classification and Regression Trees:

In this section, we delve into the utilisation of decision tree methodology, which enables non-linear classification and offers transparent and interpretable decision rules. Furthermore, we explore the Random Forest approach, which addresses the limitations of decision trees by introducing randomness to enhance model accuracy and mitigate potential issues related to overfitting.

i.) Decision Tree:

The decision tree method involves selecting a complexity parameter (cp) that serves as a pruning threshold to decide whether to prune a branch in the tree. In the presented Figure 4, the cp value of 0.003262 is obtained by choosing the maximum ROC performance metric. As depicted in Figure 4, the cp value of 0.003262 indicates that a moderate level of pruning is applied in the decision tree model.



```
CART

8582 samples
  22 predictor
   2 classes: 'neutralordissatisfied', 'satisfied'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 6866, 6866, 6866, 6866, 6864, 6865, ...
Resampling results across tuning parameters:

  cp           ROC        Sens       Spec
  0.003262643  0.9529306  0.9230910  0.9220101
  0.003845258  0.9510635  0.9186656  0.9251931
  0.005826148  0.9464867  0.9063151  0.9320284
  0.006059194  0.9465645  0.9064703  0.9314076
  0.010331702  0.9198831  0.8930317  0.9067791
  0.011652296  0.9082505  0.8835518  0.9018893
  0.020508040  0.8974165  0.8648323  0.8990124
  0.047541366  0.8759110  0.8775720  0.8505413
  0.091353997  0.8351428  0.8221267  0.8395061
  0.618503845  0.6814311  0.8487778  0.5140844

ROC was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.003262643.
```

Figure 4: Decision Tree Model Output (CP Value)

Figure 5 visually represents the decision tree obtained from the optimal model, using the identified cp value. It is noteworthy that the root node of the decision tree is the "onlineboarding" feature. Based on this feature, the tree branches out to other features such as "Inflightwifiservice," "Type of Travel," and "Class," leading to the formation of leaf nodes representing different levels of customer satisfaction for the airline company.

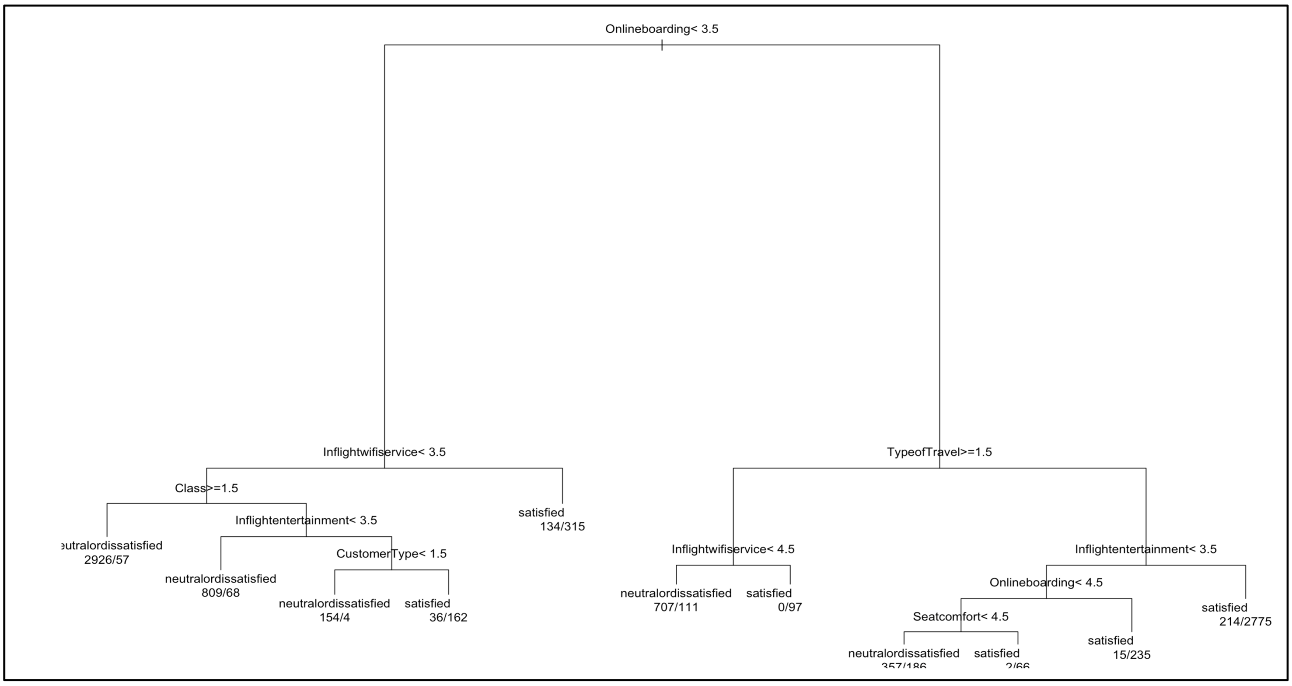


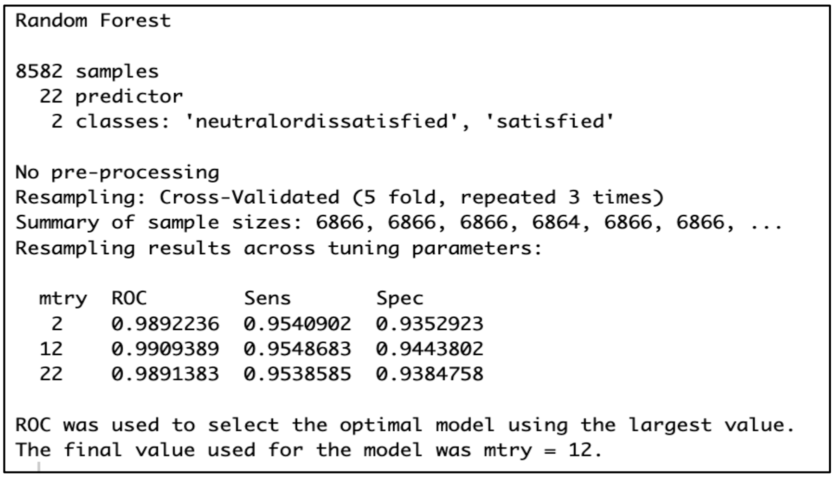
Figure 5: Obtained Decision Tree for Customer Satisfaction

In summary, the obtained cp value and the corresponding decision tree structure provide insights into the important features and their relationships in predicting customer satisfaction.

ii.) Random Forest:

Random Forest is a technique used to overcome the fundamental flaw of the decision tree where nodes are selected at random. In this method mtry parameter is obtained as shown in the Figure

6 (mtry = 12), to decide the number of variables to be included in pool from which nodes are randomly selected. Based on this mtry value, model is built to predict customer satisfaction.



```
Random Forest

8582 samples
  22 predictor
   2 classes: 'neutralordissatisfied', 'satisfied'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 6866, 6866, 6866, 6864, 6866, 6866, ...
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
   2    0.9892236  0.9540902  0.9352923
  12    0.9909389  0.9548683  0.9443802
  22    0.9891383  0.9538585  0.9384758

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 12.
```

Figure 6: Output of Random Forest

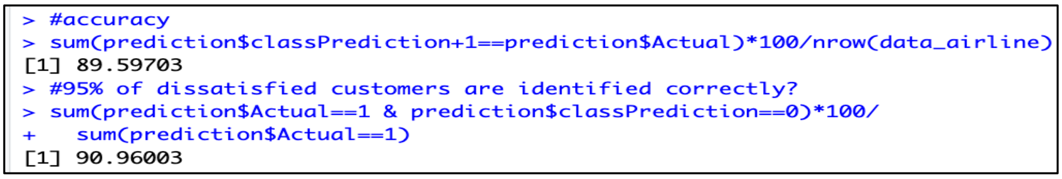
MODEL SELECTION:

The Performance metrics of the above-mentioned model is determined using the accuracy, specificity, and sensitivity. The below table shows the performance matrix for each classification method used.

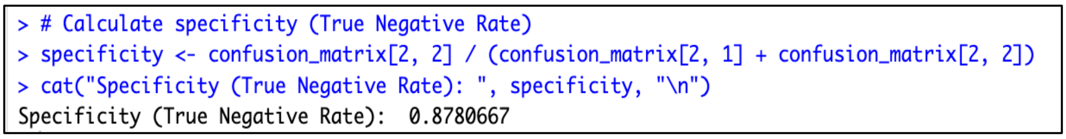
Classification Method	**Sensitivity**	**Specificity**	**Accuracy**
Logistic Regression	90.9%	87.8%	89.5%
KNN	93.7%	89.0%	91.6%
Decision Tree	93.4%	89.5%	91.7%
Random Forest	**95.4%**	**92.9%**	**94.3%**

Table 1: Comparison of Performance metrics of Classification Model

Based on the airline manager objective of correctly identifying 95% of dissatisfied customers (Sensitivity) and no more than 10% of satisfied customers are mistakenly predicted to be dissatisfied (Specificity), it can be inferred from the table that random forest model satisfies the objective. Moreover, the accuracy of the model seems to be higher compared to other models. The performance metrics output obtained from each of the model is represented in the below figures 7,8,9 and 10.

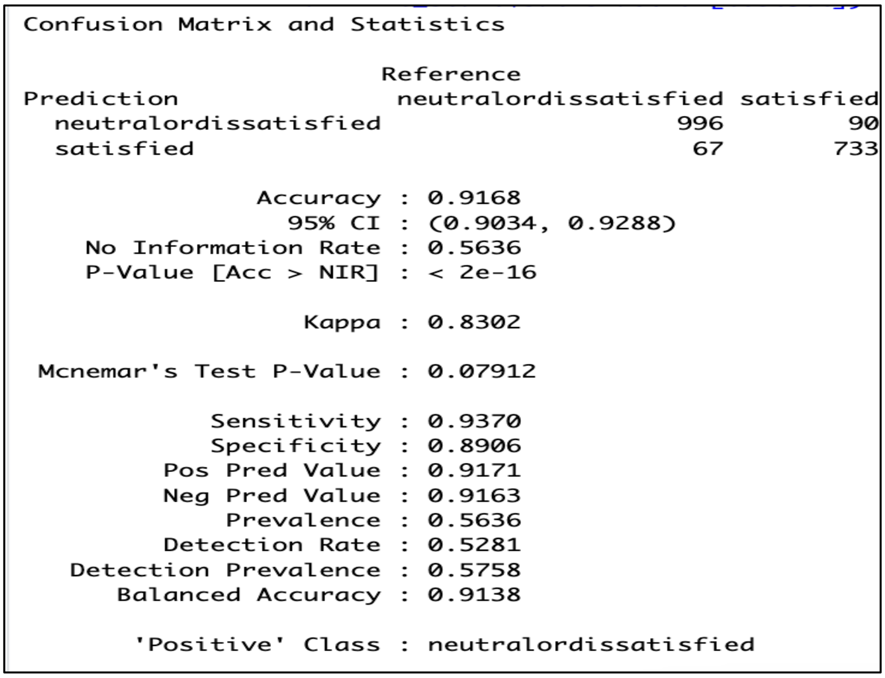


```
> #accuracy
> sum(prediction$classPrediction+1==prediction$Actual)*100/nrow(data_airline)
[1] 89.59703
> #95% of dissatisfied customers are identified correctly?
> sum(prediction$Actual==1 & prediction$classPrediction==0)*100/
+    sum(prediction$Actual==1)
[1] 90.96003
```



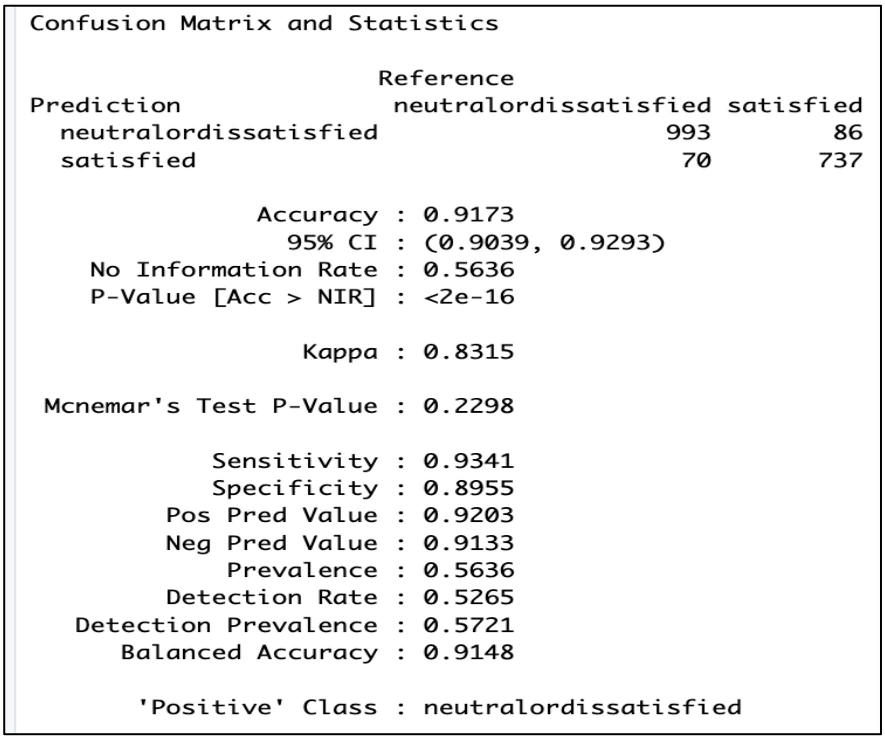
```
> # Calculate specificity (True Negative Rate)
> specificity <- confusion_matrix[2, 2] / (confusion_matrix[2, 1] + confusion_matrix[2, 2])
> cat("Specificity (True Negative Rate): ", specificity, "\n")
Specificity (True Negative Rate):  0.8780667
```

Figure 7: Performance metrics of Logit Model



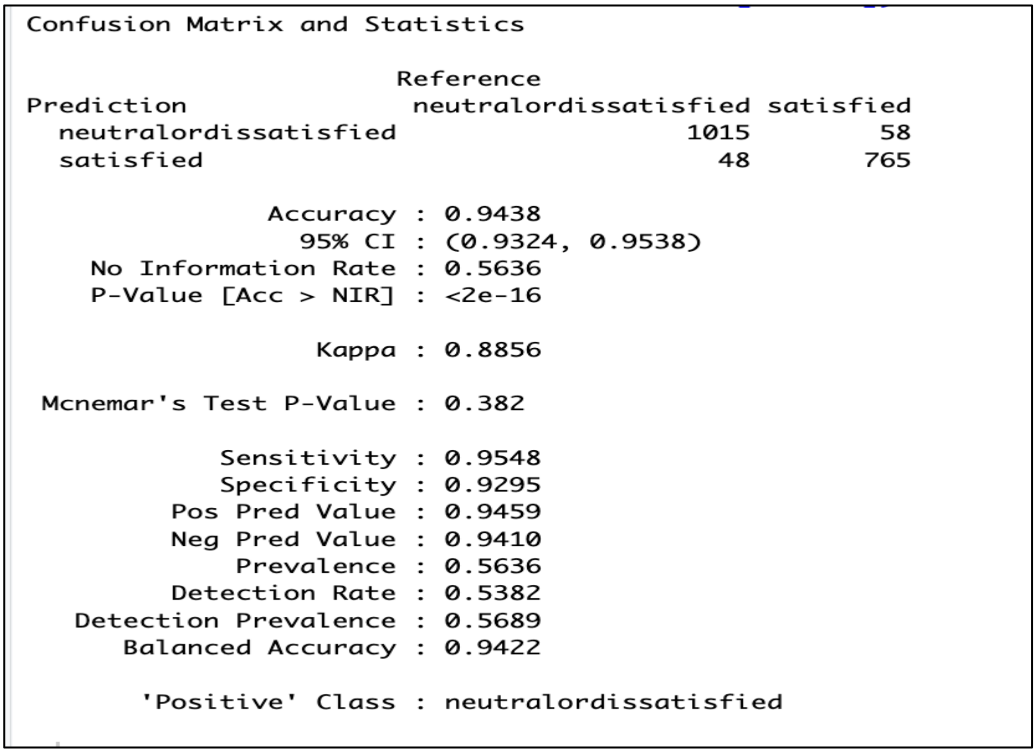
```
Confusion Matrix and Statistics

                       Reference
Prediction              neutralordissatisfied satisfied
  neutralordissatisfied                   996        90
  satisfied                                67       733
                                                        
               Accuracy : 0.9168          
                 95% CI : (0.9034, 0.9288)
    No Information Rate : 0.5636          
    P-Value [Acc > NIR] : < 2e-16         
                                          
                  Kappa : 0.8302          
                                          
 Mcnemar's Test P-Value : 0.07912         
                                          
            Sensitivity : 0.9370          
            Specificity : 0.8906          
         Pos Pred Value : 0.9171          
         Neg Pred Value : 0.9163          
             Prevalence : 0.5636          
         Detection Rate : 0.5281          
   Detection Prevalence : 0.5758          
      Balanced Accuracy : 0.9138          
                                          
       'Positive' Class : neutralordissatisfied
```

Figure 8: Performance Metrics of KNN Model



```
Confusion Matrix and Statistics

                       Reference
Prediction              neutralordissatisfied satisfied
  neutralordissatisfied                   993        86
  satisfied                                70       737
                                                        
               Accuracy : 0.9173          
                 95% CI : (0.9039, 0.9293)
    No Information Rate : 0.5636          
    P-Value [Acc > NIR] : <2e-16          
                                          
                  Kappa : 0.8315          
                                          
 Mcnemar's Test P-Value : 0.2298          
                                          
            Sensitivity : 0.9341          
            Specificity : 0.8955          
         Pos Pred Value : 0.9203          
         Neg Pred Value : 0.9133          
             Prevalence : 0.5636          
         Detection Rate : 0.5265          
   Detection Prevalence : 0.5721          
      Balanced Accuracy : 0.9148          
                                          
       'Positive' Class : neutralordissatisfied
```

Figure 9: Performance Metrics of Decision Tree Model



```
Confusion Matrix and Statistics

                       Reference
Prediction              neutralordissatisfied satisfied
  neutralordissatisfied                  1015        58
  satisfied                                48       765

               Accuracy : 0.9438
                 95% CI : (0.9324, 0.9538)
    No Information Rate : 0.5636
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8856

 Mcnemar's Test P-Value : 0.382

            Sensitivity : 0.9548
            Specificity : 0.9295
         Pos Pred Value : 0.9459
         Neg Pred Value : 0.9410
             Prevalence : 0.5636
         Detection Rate : 0.5382
   Detection Prevalence : 0.5689
      Balanced Accuracy : 0.9422

       'Positive' Class : neutralordissatisfied
```

Figure 10: Performance Metrics of Random Forest Model

TRADE-OFF of Customer Satisfaction:

The ROC curve visually represents the trade-off between correctly identifying client satisfaction and dissatisfaction in a classification model. It shows the model's performance at different threshold values, allowing decision-makers to select the optimal threshold based on their specific objectives. The Figure 11 below represents ROC Curve of Random Forest with different thresholds.

In this case, the AUC value is 0.991, indicating excellent performance of the classification model in distinguishing between satisfied and dissatisfied customers. For a threshold of 0.5, the model achieves a sensitivity of 0.954 (95.4%) and a specificity of 0.936 (93.6%) demonstrating a high rate of correctly identifying both satisfied and dissatisfied customers.

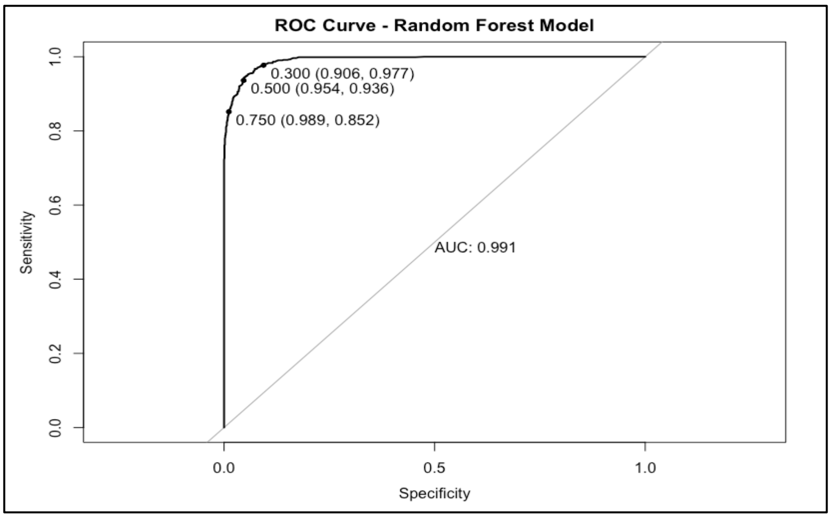


Figure 11: ROC Curve - Random Forest

By adjusting the threshold, the classification results can change. Lowering the threshold to 0.3, increases the sensitivity to 0.906 (90.6%), meaning a higher rate of correctly identifying dissatisfied customers. However, the specificity decreases to 0.977 (97.7%), resulting in a higher rate of false positives. Conversely, when the threshold is raised to 0.75, the sensitivity remains high at 0.989 (98.9%), indicating a high rate of correctly identifying satisfied customers. However, the specificity decreases to 0.852 (85.2%), leading to a higher rate of false negatives for dissatisfied customers.

The results of the classification would change accordingly based on the chosen threshold and the trade-off between correctly identifying satisfaction and dissatisfaction.

Variable Importance:

The importance of each feature in predicting customer satisfaction is assessed using the VarImp () function. This function is applied to the airline data, specifically to the random forest model, which yielded better results compared to other models. The output of the VarImp () function is presented in Figure 12, showcasing the significance of each variable in the prediction process.

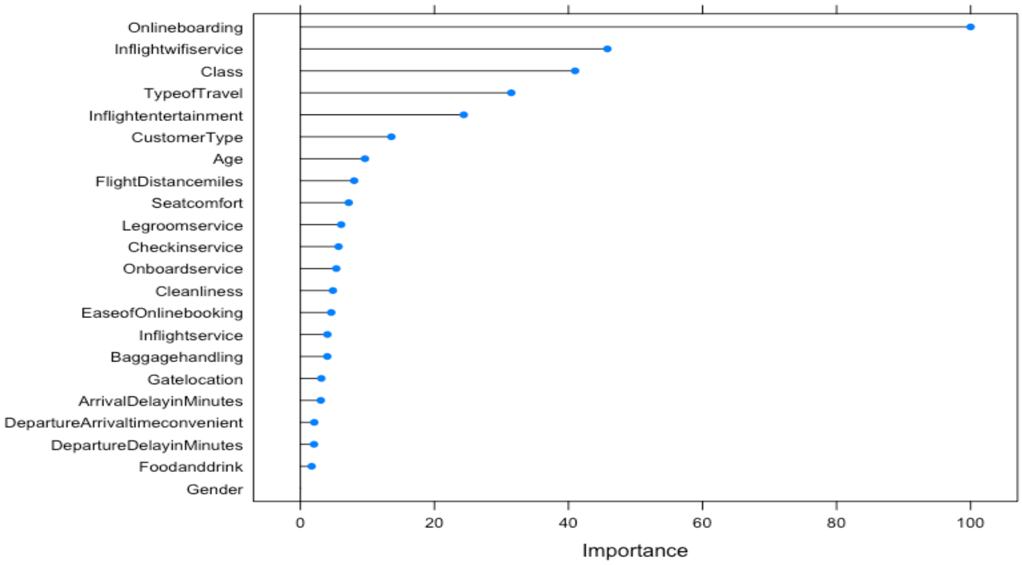
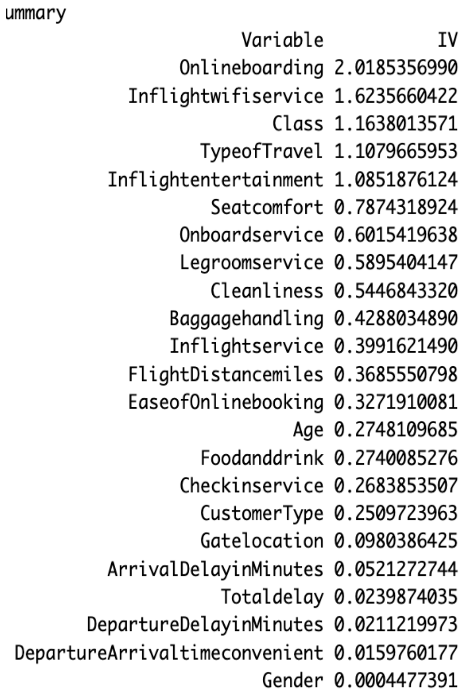


Figure 12: Variable Importance based on Random Forest

From the figure above, it can be said that Onlineboarding, Inflightwifiservice, Class, TypeofTravel, Inflightentertainment, Customer Type can be considered as important variable which might give insight on the contribution towards customer satisfaction.



```
ummary
                              Variable           IV
                        Onlineboarding 2.0185356990
                   Inflightwifiservice 1.6235660422
                                 Class 1.1638013571
                          TypeofTravel 1.1079665953
                 Inflightentertainment 1.0851876124
                           Seatcomfort 0.7874318924
                        Onboardservice 0.6015419638
                        Legroomservice 0.5895404147
                           Cleanliness 0.5446843320
                       Baggagehandling 0.4288034890
                       Inflightservice 0.3991621490
                   FlightDistancemiles 0.3685550798
                   EaseofOnlinebooking 0.3271910081
                                   Age 0.2748109685
                          Foodanddrink 0.2740085276
                        Checkinservice 0.2683853507
                          CustomerType 0.2509723963
                          Gatelocation 0.0980386425
                 ArrivalDelayinMinutes 0.0521272744
                            Totaldelay 0.0239874035
               DepartureDelayinMinutes 0.0211219973
       DepartureArrivaltimeconvenient 0.0159760177
                                Gender 0.0004477391
```

Figure 13: IV Values

These variables align with the Information Value (IV) values (Refer Figure 13), which were previously identified as important in the data analysis.

The variables Type of Travel and Class were found to play a major role in attracting clients of different segments, and there were varying responses related to onlineboarding, inflightwifiservice, and inflight entertainment, suggesting that these variables may also play an important role in customer satisfaction which was discussed in detail in the initial data analysis. Thus, these variables are considered important for predicting customer satisfaction.

Generalisation Performance of Models:

The ROC curves and AUC values (Refer Figure 14) provide valuable insights into the expected generalization performance of different models when deployed.

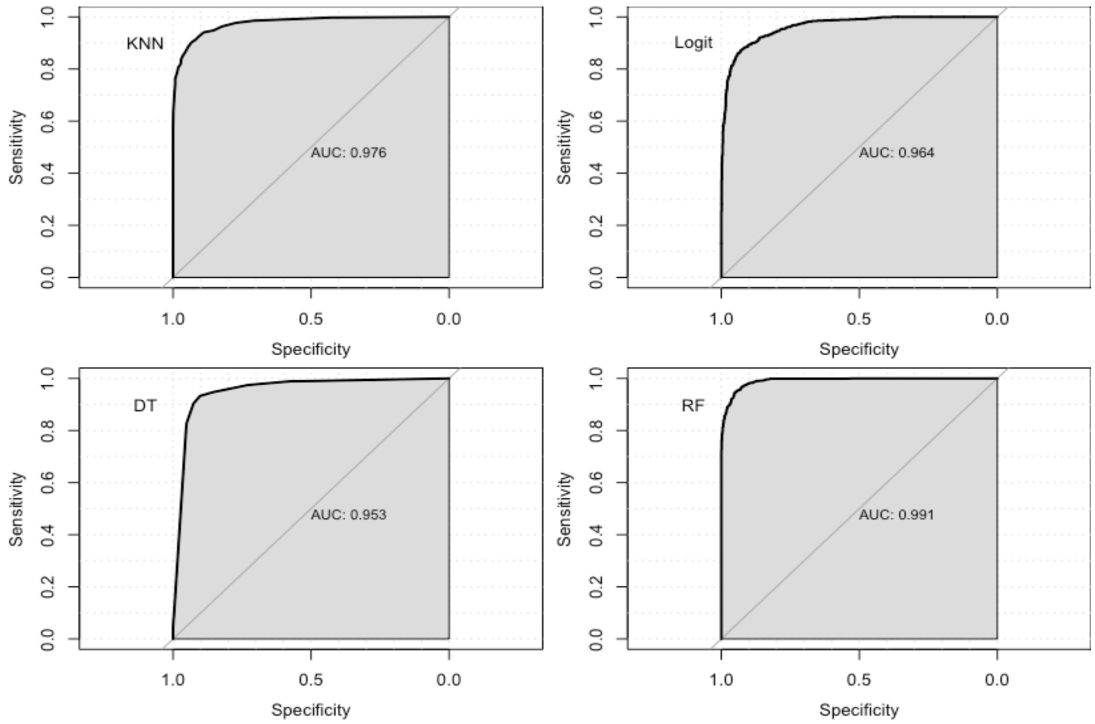


Figure 14: Generalisation ROC Curves for Classification Methods

The KNN model, with an AUC of 0.976, demonstrates strong performance in real-world scenarios and its ability to reliably predict customer satisfaction. Similarly, the Logit model, with an AUC of 0.964, is also expected to perform well, although slightly lower than the KNN model, in providing meaningful predictions for customer satisfaction. Comparatively, the Decision Tree model exhibits a lower AUC of 0.953, indicating a relatively reduced generalization performance. However, it remains a useful tool for predicting customer satisfaction, albeit with slightly lower accuracy compared to other models.

In contrast, the Random Forest (RF) model stands out with the highest AUC of 0.991, highlighting its excellent discrimination ability. By combining multiple decision trees and performing feature selection, the RF model delivers robust and accurate predictions. As a result, it is anticipated to exhibit superior generalization performance when deployed, making it a compelling choice for predicting customer satisfaction.

To summarize, based on the AUC values, the Random Forest model is expected to demonstrate the highest generalization performance, followed by the KNN and Logit models. These models, when deployed, possess the potential to accurately identify customer satisfaction and effectively support decision-making processes within the airline industry

CONCLUSION:

In summary, the analysis aimed to predict customer satisfaction using four classification models: Logistic Regression, KNN, Decision Tree, and Random Forest. The Random Forest model outperformed the others, with an impressive AUC of 0.991, indicating its superior generalization performance. The key variables influencing customer satisfaction included Onlineboarding, Inflightwifiservice, Class, TypeofTravel, Inflightentertainment, and Customer Type. The ROC curve provided valuable insights into the trade-off between correctly identifying satisfied and dissatisfied customers, allowing for threshold adjustments based on specific objectives.

Overall, the Random Forest model was deemed the most appropriate for accurately predicting customer satisfaction and informing decision-making in the airline industry due to its strong performance and capacity to handle complex relationships. These findings empower airlines to gain deeper insights into customer satisfaction drivers and make data-informed choices to improve the overall customer experience.