MSCI Coursework Part 2

Rivyesch Ranjan – 36330520

(Count – 2099 words)

Executive Summary

Classification was done on data "AirlinesData81" to determine if a customer is satisfied or not. Several common methods such as logistic regression, KNN and decision trees were explored. For each method a few candidate models were built on different variable subsets of the dataset. Furthermore, the effect of the threshold selected on the prediction of the classifier is examined. It was found that at low thresholds sensitivity is higher whilst at high thresholds specificity is higher. Another finding was that sensitivity and specificity are inversely proportional, thus there is always a trade-off. The best model for each method was unable to satisfy both the marketing manager's requirement at the same time. However, techniques that improve on the flaws of decision tree models such as bagging and Random Forest were successful in making accurate predictions that satisfy both sensitivity and specificity benchmarks. These two models should give reliable predictions when deployed on new data since they also have high testing accuracy of over 94%. The study also shed light on the most important variables for the accurate prediction of satisfaction and dissatisfaction. Features online boarding, inflight wifi service, class, type of travel, inflight entertainment and leg room are considered essential in the training and development of accurate models. Other features that are useful are onboard service, cleanliness and flight distance.

1 Introduction

The UK airline intends to make use of the data and findings from the previous study to obtain an appropriate classification model that can make predictions on the satisfaction or dissatisfaction of their customers. This study is an extension on the EDA performed in the previous study. The aim is to develop several common classifier models and compare the approaches to identify the one that best allows the airline company to make correct decisions. The marketing manager has outlined that a classification model makes the correction decision if it's prediction can meet the following criteria:

A. 95% of dissatisfied customers are identified correctly
B. At most 10% of satisfied customers are mistakenly predicted to be dissatisfied

The structure of the report is split into five main sections. Methodology covers the choice of variable subsets used to train the classifiers along with the criteria for selecting the best classification model for each method. The next three section details the model building process, results of the candidate models built and the best performing model for logistic regression, k-Nearest Neighbour (KNN) and decision trees. The section on decision trees also covers bagging and Random Forest. The final main section covers the performance evaluation of the best models of each classification method applied and provides a final recommendation on the most suitable model.

2 Methodology

When dealing with Gradient Descent or Distance Based algorithms such as logistic regression and KNN that are very sensitive to the range of the data points it becomes necessary to

perform feature scaling. This ensures that the variables are on a similar scale and have equal influence in the model. Only numerical features were scaled using the min-max approach. The categorical and ordinal features were converted to dummy variables.

Based on the previous study features that were poor predictor variables based on EDA, low IV and/or high multicollinearity were filtered out from the dataset. This is considered as variable subset 1. From subset 1 two other variable subsets are created. Subset 2 consists of only the important variables based on the MDS plots in the previous study. Subset 3 are the variables that are selected by Elastic Net regularization with α=0.6. The coefficients of the variables that cannot be shrunk down to zero and hence are important are shown below. More details on the regularization process performed can be found in Section 3.3 of the previous study.

	s1		
(Intercept)	-1.430735e+00	Seat.comfort.Q	7.760790e-01
cleanData.Age	.	Seat.comfort.C	1.015451e-03
cleanData.Flight.Distance	1.393191e-05	Seat.comfort.4	-2.069134e-01
cleanData.Departure.Delay.in.Minutes	.	Inflight.entertainment.L	3.069424e-01
cleanData.Arrival.Delay.in.Minutes	-3.693295e-03	Inflight.entertainment.Q	-8.251818e-01
GenderMale	.	Inflight.entertainment.C	-8.060443e-02
Customer.TypeLoyal.Customer	2.634696e+00	Inflight.entertainment.4	.
Type.of.TravelPersonal.Travel	-3.018750e+00	On.board.service.L	1.061262e+00
ClassEco	-8.742660e-01	On.board.service.Q	.
ClassEco.Plus	-8.769470e-01	On.board.service.C	.
Inflight.wifi.service.L	3.072282e+00	On.board.service.4	4.964623e-03
Inflight.wifi.service.Q	1.993979e+00	Leg.room.service.L	7.644952e-01
Inflight.wifi.service.C	1.562650e-01	Leg.room.service.Q	1.143887e-01
Inflight.wifi.service.4	.	Leg.room.service.C	.
Departure.Arrival.time.convenient.L	-8.310276e-01	Leg.room.service.4	-2.187413e-01
Departure.Arrival.time.convenient.Q	-2.065457e-01	Baggage.handling.L	3.699102e-01
Departure.Arrival.time.convenient.C	6.448128e-02	Baggage.handling.Q	5.023715e-01
Departure.Arrival.time.convenient.4	2.291564e-02	Baggage.handling.C	2.347246e-01
Ease.of.Online.booking.L	6.264063e-01	Baggage.handling.4	-2.077595e-01
Ease.of.Online.booking.Q	.	Checkin.service.L	6.172844e-01
Ease.of.Online.booking.C	.	Checkin.service.Q	1.160356e-01
Ease.of.Online.booking.4	.	Checkin.service.C	1.999461e-01
Gate.location.L	-4.725227e-01	Checkin.service.4	.
Gate.location.Q	-1.255652e-01	Inflight.service.L	1.417261e-01
Gate.location.C	9.861708e-02	Inflight.service.Q	7.213042e-01
Gate.location.4	.	Inflight.service.C	.
Food.and.drink.L	.	Inflight.service.4	-3.437698e-01
Food.and.drink.Q	.	Cleanliness.L	6.298649e-01
Food.and.drink.C	.	Cleanliness.Q	.
Food.and.drink.4	.	Cleanliness.C	.
Online.boarding.L	1.970021e+00	Cleanliness.4	2.105342e-01
Online.boarding.Q	1.135313e+00		
Online.boarding.C	.		
Online.boarding.4	-5.297474e-01		
Seat.comfort.L	.		

Figure 1: Elastic Net Coefficients

Another method called recursive feature elimination (RFE) which applies a backward selection process to find the optimal combination of features is also used to create Subset 4. The rfe function performs a 10-fold cross-validation with 5 repeats to improve the performance of the feature selection process. RFE identifies that only 15 of the 17 features in subset 1 is important. The selection is based on the highest accuracy score and the most important variables are shown in the figures below.

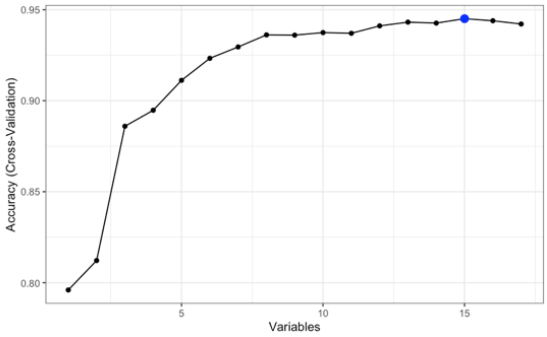

Figure 2: Optimal Number of Variables based on RFE

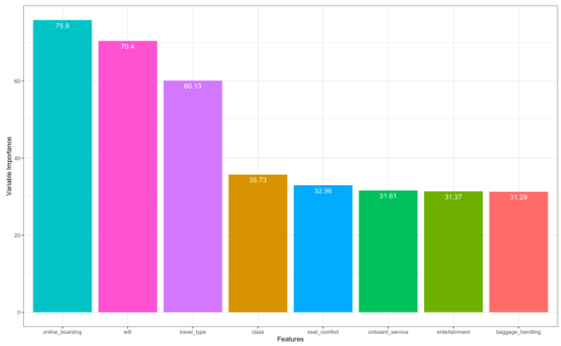

Figure 3: Most Important Variables based on RFE

Every algorithm explored in this study will be trained and tested on the variable subsets detailed below. For each algorithm a manual back-wise selection is performed starting with the variables contained in subset 1. Decisions on variables removed iteratively are based on statistical significance of the predictors or variable importance to the model.

Table 1: Selection of Variables for Each Subset

Variable	Subset 1	Subset 2	Subset 3	Subset 4
Age	Y	N	N	Y
Travel type	Y	Y	Y	Y
Class	Y	Y	Y	Y
Flight distance	Y	Y	Y	Y
Wifi service	Y	Y	Y	Y
Online booking	Y	N	Y	Y
Food drink	Y	Y	N	N
Online boarding	Y	Y	Y	Y
Seat comfort	Y	Y	Y	Y
Entertainment	Y	Y	Y	Y
Onboard service	Y	Y	Y	Y
Leg room	Y	Y	Y	Y
Baggage handling	Y	Y	Y	Y
Check-in service	Y	N	Y	Y
Inflight service	Y	Y	Y	Y
Cleanliness	Y	Y	Y	Y
Arrival delay	Y	N	Y	N
Gender	N	N	N	N
Departure/Arrival Time Convenient	N	N	N	N
Gate Location	N	N	N	N
Departure Delay in Minutes	N	N	N	N
Customer Type	N	N	N	N

The dataset is randomly split into training (80%) and test (20%) datasets. The choice of the best candidate model selected for each classification method is based on the test accuracy, sensitivity and specificity. Criteria A mentioned above refers to the true negative rate or

specificity which should be greater than 95%. Meanwhile criteria B can be inferred as the true positive rate or sensitivity should be greater than 90%.

3 Logistic Regression

A logistic model was created for each of the four variable subsets on the training dataset. The AIC of each model along with the training accuracy is determined. However, it is the testing accuracy that indicates the suitability of each model for the classification problem at hand. Initially, each models' prediction is made using an arbitrary threshold of 0.5.

Manual backward selection is also performed. A logistic model with all variables in subset 1 is fit and the p-value of the coefficients are investigated. When all dummy variables of a particular variable have a p-value greater than 0.05 it is inferred that the variable is statistically insignificant as a predictor and can be removed. This is done one variable at a time and repeated until not a single variable has a p-value greater than 0.05 for all of its dummy variables. As a result, a new subset 5 with food_drink and online_booking removed from subset 1 is created.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.538482	0.184740	2.915	0.00356	**
age	0.698012	0.264028	2.644	0.00820	**
travel_typePersonal Travel	-2.717708	0.152532	-17.817	< 2e-16	***
classEco	-1.358347	0.126092	-10.773	< 2e-16	***
classEco Plus	-1.114622	0.205252	-5.431	5.62e-08	***
flight_dist	1.649513	0.245960	6.706	1.99e-11	***
wifi.L	4.621077	0.328415	14.071	< 2e-16	***
wifi.Q	3.517864	0.286973	12.259	< 2e-16	***
wifi.C	1.221798	0.193560	6.312	2.75e-10	***
wifi^4	0.257274	0.132946	1.935	0.05297	.
online_booking.L	0.037099	0.152585	0.243	0.80790	
online_booking.Q	-0.085594	0.142380	-0.601	0.54773	
online_booking.C	0.221257	0.126939	1.743	0.08133	.
online_booking^4	0.127569	0.121672	1.048	0.29442	
food_drink.L	-0.301209	0.158198	-1.904	0.05691	.
food_drink.Q	-0.208998	0.145536	-1.436	0.15099	
food_drink.C	0.087523	0.132573	0.660	0.50913	
food_drink^4	0.138126	0.127976	1.079	0.28045	
online_boarding.L	2.581186	0.160219	16.110	< 2e-16	***
online_boarding.Q	1.264769	0.142220	8.893	< 2e-16	***
online_boarding.C	-0.268711	0.122533	-2.193	0.02831	*
online_boarding^4	-0.628785	0.113491	-5.540	3.02e-08	***
seat_comfort.L	-0.296927	0.165490	-1.794	0.07278	.
seat_comfort.Q	0.886360	0.147133	6.024	1.70e-09	***
seat_comfort.C	0.225293	0.141542	1.592	0.11145	
seat_comfort^4	-0.609944	0.128185	-4.758	1.95e-06	***
entertainment.L	1.368608	0.220429	6.209	5.34e-10	***
entertainment.Q	-1.332097	0.204924	-6.500	8.01e-11	***
entertainment.C	-0.571357	0.185030	-3.088	0.00202	**
entertainment^4	0.068453	0.171248	0.400	0.68935	
onboard_service.L	0.901113	0.159400	5.653	1.58e-08	***
onboard_service.Q	0.174418	0.136131	1.281	0.20011	
onboard_service.C	-0.095162	0.132611	-0.718	0.47300	
onboard_service^4	0.009475	0.120102	0.079	0.93712	
leg_room.L	0.858863	0.135153	6.355	2.09e-10	***
leg_room.Q	0.279607	0.128052	2.184	0.02900	*
leg_room.C	-0.031660	0.109736	-0.289	0.77296	
leg_room^4	-0.274628	0.106948	-2.568	0.01023	*
baggage_handling.L	-0.104977	0.167843	-0.625	0.53168	
baggage_handling.Q	0.657756	0.146259	4.497	6.88e-06	***
baggage_handling.C	0.562095	0.135423	4.151	3.32e-05	***
baggage_handling^4	-0.415190	0.126266	-3.288	0.00101	**
checkin_service.L	0.505610	0.127295	3.972	7.13e-05	***
checkin_service.Q	0.128882	0.108755	1.185	0.23599	
checkin_service.C	0.320332	0.110378	2.902	0.00371	**
checkin_service^4	0.002175	0.098774	0.022	0.98243	
inflight_service.L	-0.117638	0.174502	-0.674	0.50023	
inflight_service.Q	0.629013	0.153672	4.093	4.25e-05	***
inflight_service.C	0.103756	0.142641	0.727	0.46699	
inflight_service^4	-0.423406	0.136540	-3.101	0.00193	**
cleanliness.L	0.479384	0.173281	2.767	0.00567	**
cleanliness.Q	0.041006	0.145548	0.282	0.77815	
cleanliness.C	-0.003516	0.147881	-0.024	0.98103	
cleanliness^4	0.319214	0.135713	2.352	0.01867	*
arrival_delay	-2.276731	0.892723	-2.550	0.01076	*

Figure 4: Coefficients of Logistic Regression for Subset 1

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   0.477527   0.182270   2.620 0.008796 **
age                           0.804371   0.259187   3.103 0.001913 **
travel_typePersonal Travel -2.757316   0.150180 -18.360  < 2e-16 ***
classEco                   -1.333405   0.125099 -10.659  < 2e-16 ***
classEco Plus              -1.067232   0.202849  -5.261 1.43e-07 ***
flight_dist                 1.662578   0.245124   6.783 1.18e-11 ***
wifi.L                      4.634033   0.309752  14.960  < 2e-16 ***
wifi.Q                      3.430203   0.265833  12.904  < 2e-16 ***
wifi.C                      1.342223   0.174968   7.671 1.70e-14 ***
wifi^4                      0.322637   0.111748   2.887 0.003887 **
online_boarding.L           2.595375   0.159228  16.300  < 2e-16 ***
online_boarding.Q           1.254338   0.141527   8.863  < 2e-16 ***
online_boarding.C          -0.223598   0.120575  -1.854 0.063677 .
online_boarding^4          -0.603751   0.111093  -5.435 5.49e-08 ***
seat_comfort.L             -0.328824   0.163306  -2.014 0.044057 *
seat_comfort.Q              0.867658   0.145945   5.945 2.76e-09 ***
seat_comfort.C              0.242148   0.139205   1.740 0.081946 .
seat_comfort^4             -0.607354   0.125666  -4.833 1.34e-06 ***
entertainment.L             1.214686   0.205683   5.906 3.51e-09 ***
entertainment.Q            -1.419419   0.190823  -7.438 1.02e-13 ***
entertainment.C            -0.574661   0.174071  -3.301 0.000962 ***
entertainment^4             0.108498   0.162508   0.668 0.504361
onboard_service.L           0.947114   0.158324   5.982 2.20e-09 ***
onboard_service.Q           0.205985   0.135361   1.522 0.128075
onboard_service.C          -0.100494   0.131640  -0.763 0.445223
onboard_service^4          -0.005263   0.119503  -0.044 0.964873
leg_room.L                  0.863778   0.134463   6.424 1.33e-10 ***
leg_room.Q                  0.295412   0.127250   2.322 0.020259 *
leg_room.C                 -0.048572   0.109214  -0.445 0.656509
leg_room^4                 -0.280435   0.106749  -2.627 0.008613 **
baggage_handling.L         -0.074716   0.166759  -0.448 0.654117
baggage_handling.Q          0.689322   0.144701   4.764 1.90e-06 ***
baggage_handling.C          0.544662   0.134210   4.058 4.94e-05 ***
baggage_handling^4         -0.418523   0.125430  -3.337 0.000848 ***
checkin_service.L           0.509527   0.126421   4.030 5.57e-05 ***
checkin_service.Q           0.135053   0.108053   1.250 0.211341
checkin_service.C           0.316741   0.109829   2.884 0.003927 **
checkin_service^4           0.009650   0.098241   0.098 0.921755
inflight_service.L         -0.097547   0.173972  -0.561 0.574998
inflight_service.Q          0.650871   0.152926   4.256 2.08e-05 ***
inflight_service.C          0.090273   0.141992   0.636 0.524933
inflight_service^4         -0.419576   0.135410  -3.099 0.001945 **
cleanliness.L               0.432700   0.168178   2.573 0.010086 *
cleanliness.Q              -0.017050   0.141113  -0.121 0.903827
cleanliness.C               0.029704   0.143131   0.208 0.835594
cleanliness^4               0.365325   0.130327   2.803 0.005061 **
arrival_delay              -2.154534   0.902697  -2.387 0.016997 *
```

Figure 5: Coefficients of Logistic Regression for Subset 5

The table below shows the findings from the experimentation using logistic regression models. “logitModel1” in the table refers to the logistic model trained on variable subset 1, and so on. The table shows that logitModel1, logitModel2 and logitModel4 resulted in the highest test accuracy. The training accuracy was generally higher than the testing accuracy. However, this is expected since the training data has been seen by the model. However, none of these models are able to meet the sensitivity and specificity criteria given by the marketing manager.

Table 2: Candidate Logistic Regression Models' Performance

Models	AIC	Train Accuracy	Test Accuracy	Sensitivity	Specificity
logitModel1	3436.5	0.920	0.914	0.883	0.938
logitModel2	3464.2	0.919	0.914	0.875	0.945
logitModel3	3444.3	0.920	0.911	0.882	0.935
logitModel4	3442.3	0.922	0.914	0.883	0.939
logitModel5	3434.3	0.920	0.912	0.882	0.936

The effect of the selected threshold was tested on the second and fourth models separately. Preliminary testing found that the testing accuracy deteriorated significantly for thresholds at either end of the range. Therefore, only thresholds from 0.3 to 0.7 are displayed in the table. At high thresholds, the prediction will get more negative values which results in a higher specificity and lower sensitivity. At low thresholds there are more positive values which leads to an increase in sensitivity and decrease in specificity. It is evident from the results that selecting a model with a higher specificity inadvertently leads to lower sensitivity. Regardless of the threshold selection no logistic regression model is capable of meeting the sensitivity and specificity requirements for this dataset. The best trained logistic model is logitModel4 since it has the highest train and test accuracy, and barely misses on both criteria.

Table 3: Effect of Prediction Threshold on Logistic Regression Models' Performance

Threshold	logitModel2				logitModel4			
	Train Accuracy	Test Accuracy	Sensitivity	Specificity	Train Accuracy	Test Accuracy	Sensitivity	Specificity
0.3	0.904	0.901	0.926	0.882	0.904	0.906	0.926	0.890
0.35	0.909	0.910	0.918	0.904	0.911	0.909	0.918	0.903
0.4	0.914	0.913	0.907	0.918	0.915	0.911	0.907	0.915
0.45	0.918	0.913	0.892	0.930	0.919	0.912	0.893	0.926
0.5	0.919	0.914	0.875	0.945	0.922	0.914	0.883	0.939
0.55	0.919	0.911	0.857	0.954	0.919	0.910	0.862	0.947
0.6	0.917	0.910	0.842	0.964	0.919	0.911	0.846	0.962
0.65	0.913	0.906	0.823	0.972	0.916	0.903	0.819	0.969
0.7	0.909	0.899	0.803	0.975	0.911	0.899	0.805	0.973

4 K-Nearest Neighbours (k-NN)

Similar to Section 3, a kNN model is trained for each of the four variable subsets. For each model cross validation based on random sampling with 5 folds, repeated 3 times for 10 values of k is performed to identify and select the optimal number of neighbours k. The optimal value of k is selected based on k value that gives the highest ROC.

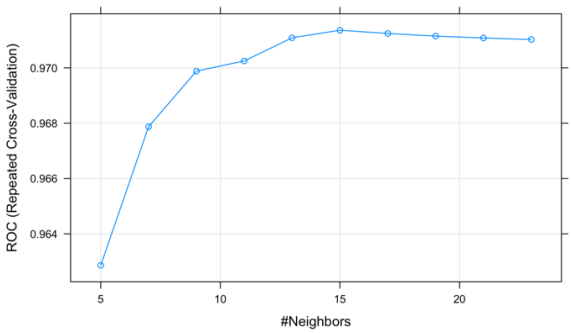

Figure 6: Optimal k neighbours for Subset 1

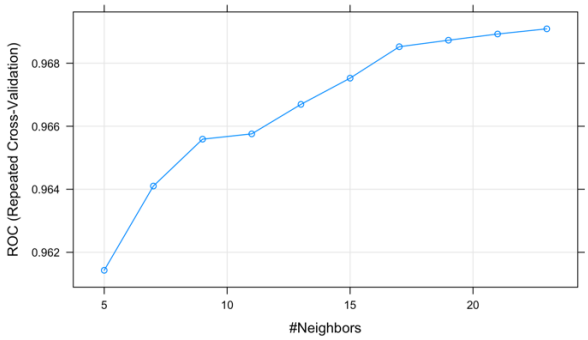

Figure 7: Optimal k neighbours for Subset 5

A manual backward selection is also performed where a variable is removed one at a time if all of its dummy variables have a variable importance below 40%. Starting with subset 1 a total of eight variables were removed to create subset 5 for which another kNN model is trained.

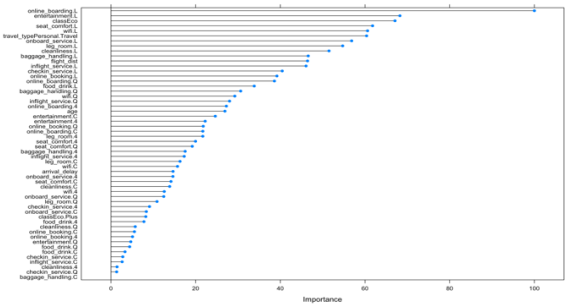

Figure 8: Variable Importance of knnTrain1 for Subset 1

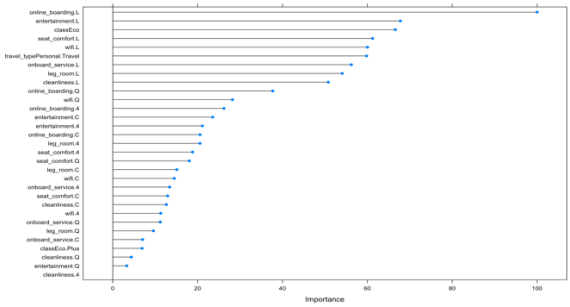

Figure 9: Variable Importance of knnTrain5 for Subset 5

Overall, the first four models trained on the first four subsets showed testing accuracy above 92%. None of these models met both criteria simultaneously, although it is evident that criteria A can be achieved. The two best candidate models are knnTrain1 and knnTrain4 since these models meets the specificity criterion and have the highest sensitivity among the other models. The results shown in the table below is for a default threshold of 0.5.

Table 4: Candidate KNN Models' Performance

Models	Optimal k	ROC	Sensitivity	Specificity	Train Accuracy	Test Accuracy
knnTrain1	15	0.971	0.884	0.954	0.928	0.923
knnTrain2	13	0.972	0.865	0.963	0.929	0.920
knnTrain3	19	0.972	0.879	0.956	0.925	0.922
knnTrain4	17	0.972	0.883	0.952	0.929	0.922
knnTrain5	23	0.969	0.8689	0.9409	0.910	0.909

The prediction threshold is varied to explore if this would lead to a prediction on the test dataset that meets both criteria. At low thresholds the sensitivity is higher and the specificity is lower, and the opposite is true for high thresholds. None of the prediction thresholds for the two candidate models are capable of meeting both criteria. The best model is knnTrain1 at a threshold of 0.5 which meets the specificity criterion and barely misses the sensitivity criterion.

Table 5: Effect of Prediction Threshold on KNN Models' Performance

Threshold	knnTrain1				knnTrain4			
	Train Accuracy	Test Accuracy	Sensitivity	Specificity	Train Accuracy	Test Accuracy	Sensitivity	Specificity
0.3	0.902	0.898	0.937	0.866	0.906	0.903	0.926	0.885
0.35	0.917	0.910	0.921	0.901	0.906	0.903	0.926	0.885
0.4	0.927	0.917	0.904	0.927	0.916	0.913	0.914	0.913
0.45	0.927	0.917	0.904	0.927	0.924	0.916	0.899	0.930
0.5	0.928	0.923	0.884	0.954	0.929	0.922	0.883	0.952
0.55	0.928	0.917	0.854	0.967	0.928	0.919	0.863	0.963
0.6	0.922	0.911	0.827	0.977	0.925	0.911	0.832	0.974
0.65	0.922	0.911	0.827	0.977	0.920	0.908	0.810	0.985
0.7	0.916	0.906	0.803	0.986	0.920	0.907	0.809	0.985

5 Decision Trees

Following the same methodology, five decision trees are fitted. The training process is similar to that of KNN except here cross validation is used to identify the optimal complexity parameter cp. Again, the best cp selected is based on the highest ROC value.

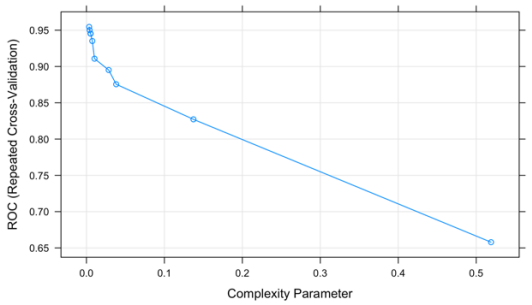

Figure 10: Optimal cp for DT1 for Subset 1

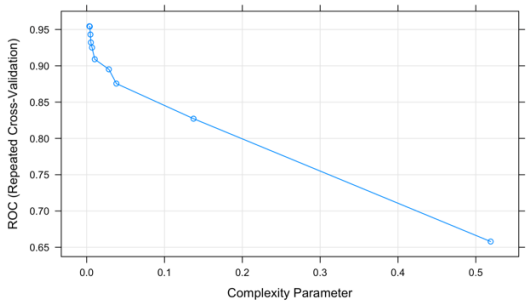

Figure 11: Optimal cp for DT5 for Subset 5

The fifth decision tree is fitted using only variables that have an importance score of 20% or higher for at least one of its dummy variables. Starting with subset 1 a total of nine variables were removed to create subset 5.

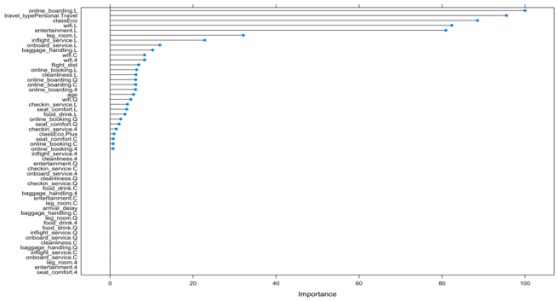

Figure 12: Variable Importance of DT1 for Subset 1

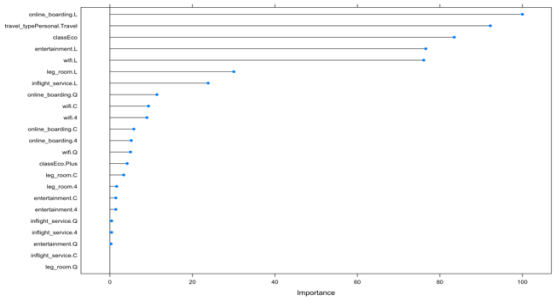

Figure 13: Variable Importance of DT5 for Subset 5

The results suggest this manual selection of variables yielded the best performing decision tree, although only the sensitivity criterion has been met. All models were very capable in predicting satisfaction instances correctly as sensitivity is high.

Table 6: Candidate Decision Tree Models' Performance

Models	Optimal cp	ROC	Sensitivity	Specificity	Train Accuracy	Test Accuracy
DT1	0.00382	0.880	0.943	0.891	0.921	0.920
DT2	0.00334	0.955	0.936	0.896	0.921	0.918
DT3	0.00382	0.953	0.943	0.891	0.921	0.920
DT4	0.00382	0.953	0.943	0.891	0.921	0.920
DT5	0.00334	0.955	0.938	0.903	0.922	0.923

The results below are for the same decision trees and variable subsets except bagging has been implemented to reduce the variance of the decision tree, and by extension, prevent overfitting. It is evident that bagging massively improves the performance of the model in predicting dissatisfaction instances as specificity is now much higher. DT3 is considered to be the best model since it has the highest test accuracy whilst needing the fewest features during training and also meeting both criteria.

Table 7: Candidate Decision Tree with Bagging Models' Performance

Models	Optimal cp	ROC	Sensitivity	Specificity	Train Accuracy	Test Accuracy
DT1	0.00382	0.983	0.906	0.959	0.999	0.936
DT2	0.00334	0.981	0.896	0.963	0.999	0.933
DT3	0.00382	0.982	0.904	0.967	0.999	0.939
DT4	0.00382	0.982	0.903	0.965	0.999	0.937
DT5	0.00334	0.982	0.903	0.965	0.999	0.937

Five random forest models are trained for the same variable subsets used above. The training process is similar to before except the largest ROC value is used to find mtry which is the optimal number of variables to include in the pool. The value of mtry is higher than the number of features due to the model considering all dummy variables created. In a similar fashion to how bagging improved the performance of decision trees, it is observed random forest which is an ensemble method that utilizes multiple trees also improves the performance. The best model selected is RF1 since it satisfies both criteria and also has the highest test accuracy.

Table 8: Candidate Random Forest Models' Performance

Models	mtry	ROC	Sensitivity	Specificity	Train Accuracy	Test Accuracy
RF1	19	0.988	0.904	0.969	1.000	0.940
RF2	11	0.986	0.886	0.973	0.999	0.935
RF3	12	0.987	0.897	0.974	1.000	0.940
RF4	12	0.987	0.898	0.968	1.000	0.937
RF5	11	0.987	0.896	0.975	1.000	0.940

6 Performance Evaluation

ROC curves are used to check and visualize the performance of the best classification model for each method. The graph below specifies how much each model is capable of distinguishing between classes at various threshold settings.

Of the five models being compared random forest models has the best performance followed by decision tree with bagging. The AUC of 0.988 and 0.981 indicate these two models are able to distinguish between satisfaction and dissatisfaction more than 98% of the time. The third best model belongs to the KNN method which gives an AUC of 0.975 which means there is a 97.5% chance that the model will be able to distinguish between the classes. Logistic regression model had an AUC of 0.968 which is slightly higher than the simple decision tree model which performed the worst. This goes to highlight the deficiencies of decision trees and its tendency to overfit.

The ROC graphs for KNN, logistic regression and decision tree also depict the inverse relationship between sensitivity and specificity. When sensitivity is increased specificity decreases, and vice versa. This effect has been observed throughout the study where it was

found for most models it was not possible to obtain a high benchmark for both sensitivity and specificity simultaneously, and any attempts to increase one inadvertently leads to the decrease in the other. If the goal of the classification was different, the results would remain mostly similar except the decision on the best candidate model would be different. Currently, the main deciding factor was meeting the sensitivity and specificity benchmarks.

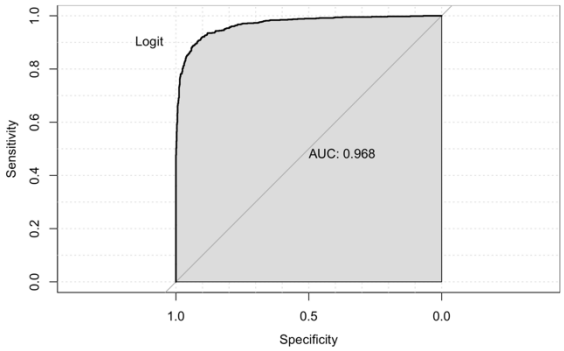

Figure 14: ROC of Best Logistic Regression Model

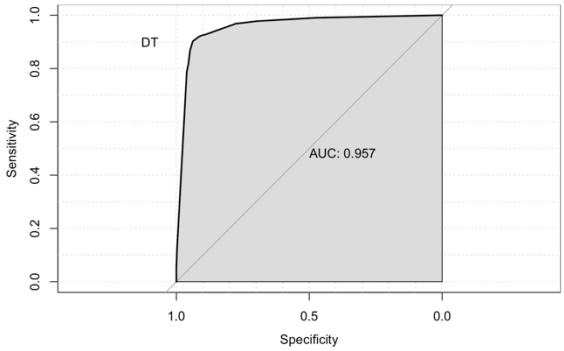

Figure 16: ROC of Best Decision Tree Model

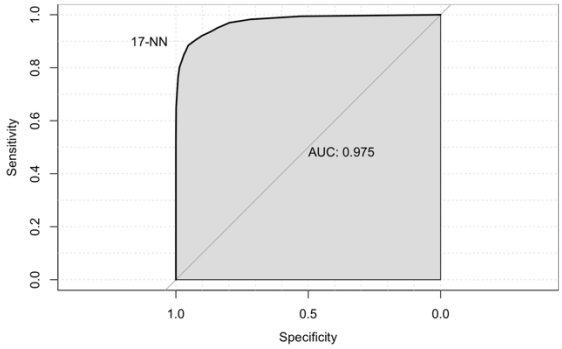

Figure 15: ROC of Best KNN Model

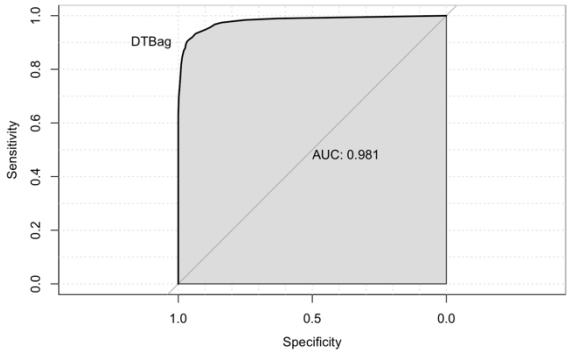

Figure 17: ROC of Best Decision Tree Model with Bagging

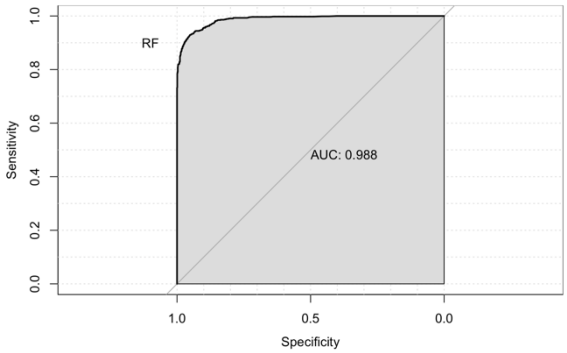

Figure 18: ROC of Best Random Forest Model

The generalization of the above models are relatively high. All models have a testing accuracy on unseen data of more than 90%. However, decision tree with bagging and Random Forest would perform the best when deployed as it has a testing accuracy of around 94%. Reliable results can be expected from these two models.

From the development of all the candidate models above, several variables were found to be mainstays in the construction of majority of the model and detailed in Table 9. Furthermore, these variables showed high scores of variable importance for KNN and Decision Tree models, whilst being statistically significant for logistic regression models. There were also several

other variables which was observed to be required by many models. These are listed out in Table 10 below.

Table 9: Most Important Variables

Feature Name
Online Boarding
Inflight Wifi Service
Class
Type of Travel
Inflight Entertainment
Leg Room

Table 10: Other Potentially Useful Variables

Feature Name
Onboard Service
Cleanliness
Flight Distance

7 Conclusion

Three main types of classification methods were explored to evaluate the effectiveness of each method for classifying the airlines data. All models had a reasonably high testing accuracy of more than 90%. However, despite training several candidate models for logistic regression, KNN and simple decision tree none of these models were good enough to make predictions that satisfy the marketing manager's requirements. It was only capable of meeting either the sensitivity criterion or the specificity criterion, and never both. Varying the threshold did not help much since sensitivity and specificity are inversely proportional to each other. Decision trees with bagging applied and random forest models, however, were able to meet both sensitivity and specificity criteria simultaneously. This goes to show that techniques to improve on decision tree do indeed improve the performance significantly and prevent overfitting through either the introduction of impurity or an ensemble of decision trees.