Prediction of Customer Satisfaction for Invistico Airlines

Hanni Chen, Joy Chow, Xuefei Qiao, Guanghui Shen

Abstract

As COVID-19 impacts the airline industry significantly, commercial airlines are suffering from losses in revenues. Customer satisfaction has become a more critical component in business recovery. This study investigates the key factors affecting customer satisfaction for Invistico Airlines by looking into customer attributes and individual ratings on boarding, in-flight, and post-flight services. It provides insights into the features that should be emphasized to improve customer experiences.

Gower+PAM and MCA+K-means methods are applied for clustering purpose in the EDA process to gain insights into customers' overall demographics and ratings of different flight services. Classification trees and logistic regression models are then built to predict customer satisfaction status. On account of accuracy and time elapsed as criteria, the boosted tree and bagged tree are selected as our best models. Entertainment and seat comfort are identified by tree related models as important services that Invistico Airlines can improve to gain a competitive edge among peers and get through difficult times when budgets are tight.

Keywords

Customer satisfaction, business, airlines, logistic regression, classification tree, bagging, random forest, boosting

1. Introduction

The development of the airline industry has changed the global economy significantly. It facilitates tourism and communications between countries economically and culturally, which help generate economic growth. The airline industry promotes improving people's living standards; therefore, besides being a means of transportation, the airline industry must also focus on customer experience. Upon the breakout of COVID-19, the airline industry has faced multiple challenges such as inflationary pressures, preventing and controlling the virus's spread while maintaining profitability. But customer satisfaction, as always been one of the airline industry's top concerns, has become more critical than ever, considering that lower customer satisfaction leads to fewer passengers and decreases in revenue. Gordin's (2013) research stated that customer dissatisfaction resulted in decades of money-losing airline operations. This also proves how vital customer satisfaction is for the recovery of air travel. According to Effler's study, attentive flight crews, flexible fares, and charges during pandemic drive record-high customer satisfaction with North American airlines. By analyzing a specific airline company, Invistico Airlines, we hope to get some valuable insights into the entire airline industry from the results.

After briefly introducing the context, we will describe our initial dataset in the next section. Then, we will perform explanatory data analysis such as multiple correspondence

analysis (MCA) for dimensionality reduction purposes and Gower's distance in Section 3. We will then list the methods used and explain the theory behind each of them in the methodology section (Section 4). Data cleaning and partitioning, demonstration of the built models, and relevant explanations are in Section 5. Finally, we will elaborate on our findings and the direction of future research.

2. Dataset

The raw dataset was obtained from Kaggle and originally used for a business competition hosted by the Indian Institute of Technology (IIT), Roorkee in 2020. It consists of over 120K observations and 23 columns that detail Invistico Airlines' customer information and their feedback on different flight services. The actual name of the airline organization is not revealed due to confidential purposes. Since the large size of the dataset may cause technical issues to our data analysis, we then randomly selected 5,000 rows from the dataset as our final sample.

Each row refers to a passenger's overall satisfaction level with the airline, along with the passenger's personal information. Specific to each customer, the dataset not only records information on flight distance, departure delay, and arrival delay, but it also employs a five-level Likert scale to collect the customer's ratings of a variety of airline services, from boarding to in-flight to post-flight. All these columns are ordinal. A "1" in the Likert scale represents "Least Satisfied" on a specific service, and a "5" means "Most Satisfied". Several ratings contain entries of 0, which stand for "not applicable" according to Kaggle description. There are also 393 missing values in the column "arrival delay in minutes".

Customer attributes and flight information are as follows:

Variable	**Description**
medi	Customer's satisfaction level with the flight (Satisfied, Dissatisfied)
Gender	Gender of the passenger (Female, Male)
Customer Type	Passenger's loyalty status (Loyal, Disloyal)
Age	Age of the passenger
Type of Travel	Purpose of the flight (Personal Travel, Business Travel)
Class	Travel class of the passenger (Business, Eco, Eco Plus)
Flight Distance	Flight distance of this journey
Departure Delay in Minutes	Minutes delayed when the flight departs
Arrival Delay in Minutes	Minutes delayed when the flight arrives

Table 1. Customer attributes and flight information

Customer ratings of services are as follows:

Variable	**Description**
Seat comfort	Passenger's satisfaction level of in-flight seat comfort
Departure/Arrival time convenient	Passenger's satisfaction level of the convenience of flight departure and arrival
Food and drink	Passenger's satisfaction level of in-flight food and drink

Gate location	Passenger's satisfaction level of gate location
In-flight wifi service	Passenger's satisfaction level of in-flight Wi-Fi service
Inflight entertainment	Passenger's satisfaction level of in-flight entertainment
Online support	Passenger's satisfaction level of online support
Ease of online booking	Passenger's satisfaction level of the convenience of online booking
On-board service	Passenger's satisfaction level of on-board service
Legroom service	Passenger's satisfaction level of in-flight legroom (space)
Baggage handling	Passenger's satisfaction level of baggage handling
Check-in service	Passenger's satisfaction level of check-in service
Cleanliness	Passenger's satisfaction level of in-flight cleanliness
Online boarding	Passenger's satisfaction level of online boarding

Table 2. Customer ratings on flight services

3. Exploratory Data Analysis

Variable "medi", which is customer satisfaction, has an entropy of 0.993, meaning the dataset is chaotic but also two classes of satisfaction are evenly distributed.

Correlation Check

Between quantitative variables, a correlation matrix in Figure 1 shows a high correlation between departure delay and arrival delay. This can cause problems in further model building, so it is wise to only keep one of these two variables when doing classification.

```
> cor(quantdata)
                                air.Age air.Flight.Distance air.Departure.Delay.in.Minutes air.Arrival.Delay.in.Minutes
air.Age                      1.00000000         -0.2713442                   -0.01534235                  -0.0169277
air.Flight.Distance         -0.27134418          1.0000000                    0.11523985                   0.1172937
air.Departure.Delay.in.Minutes -0.01534235          0.1152399                    1.00000000                   0.9742731
air.Arrival.Delay.in.Minutes   -0.01692770          0.1172937                    0.97427308                   1.0000000
```

Figure 1. Correlation Matrix

Between quantitative variables and qualitative variables, a series of spinograms were created and showed no serious correlation problem. Figures 2-5 are a few selected graphs.

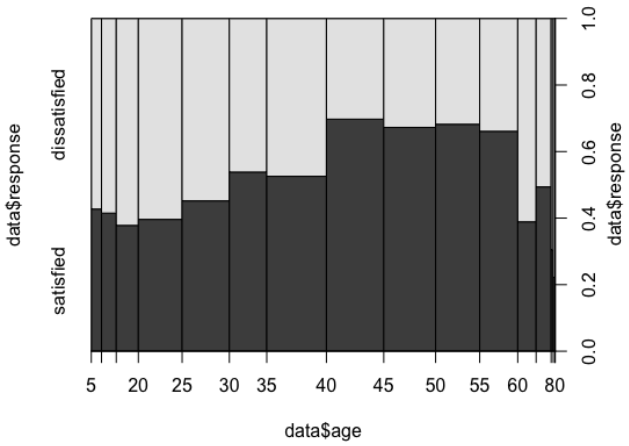

Figure 2. Spinogram for Age & Response

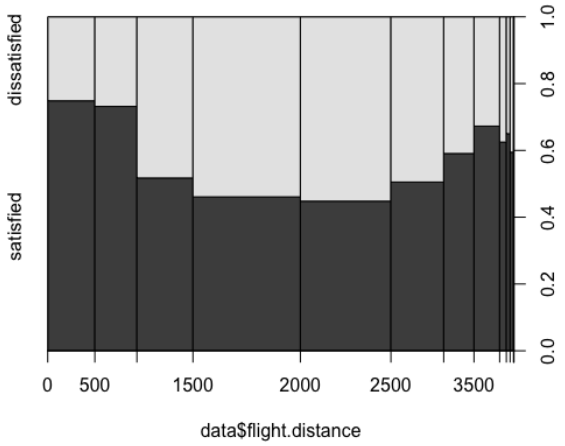

Figure 3. Spinogram for Distance & Response

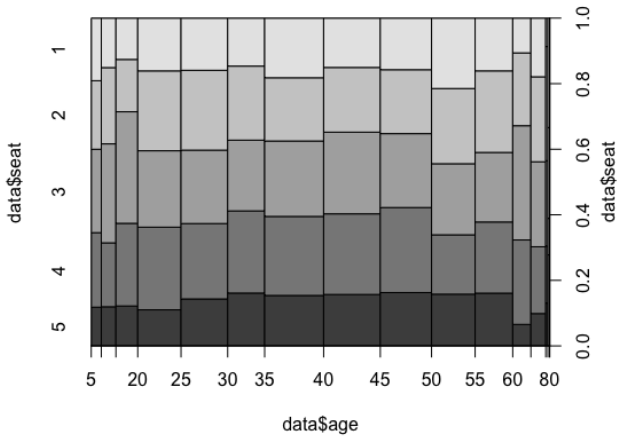

Figure 4. Spinogram for Age & Seat

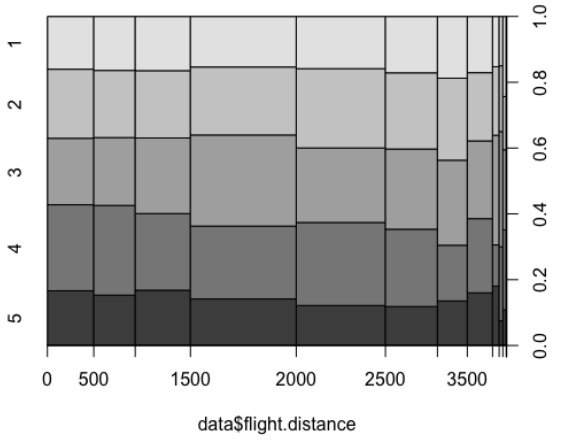

Figure 5. Spinogram for Distance & Seat

After converting passenger information and service ratings into categorical variables, we conducted clustering to separate customers into groups.

Gower+PAM

Because the features in the dataset were both quantitative and qualitative, we used Gower's distance method to measure the dissimilarity among observations. According to Figure 6, the level of separation is maximized when the number of clusters is 4, indicating that customers should be divided into 4 groups.

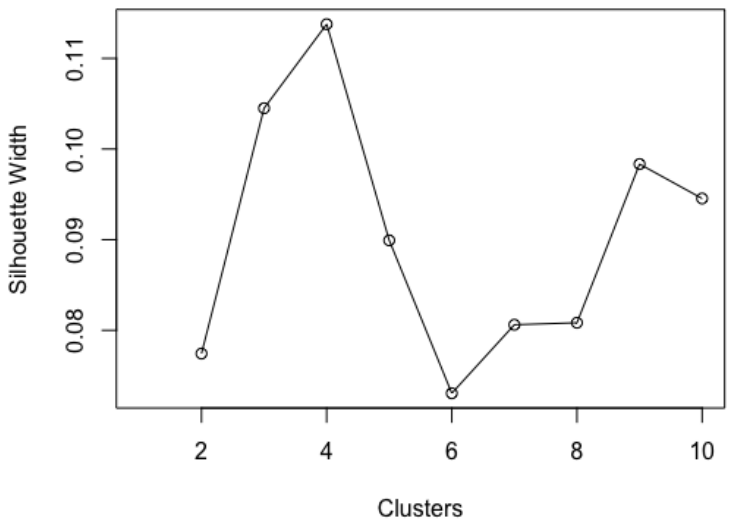

Figure 6. Clustering Diagram Using Gower's Distance

Next, we used the Partitions Around Medoids (PAM) algorithm to uncover the size as well as the representative data point of each cluster. According to Figure 7, the clusters are of sizes 1008, 1313, 891, and 1788.

```
> pam_air=pam(transformed.air, diss = TRUE, k = 4)
>
> pam_air$clusinfo
     size  max_diss   av_diss  diameter separation
[1,] 1008 0.6887768 0.4198131 0.8833917 0.05134736
[2,] 1313 0.6581866 0.3700903 0.8634463 0.01269187
[3,]  891 0.6637770 0.4350106 0.8738565 0.04960332
[4,] 1788 0.6559904 0.3802362 0.8653403 0.01269187
```

Figure 7. Clusters Separated Using PAM

Figure 8 shows that the representative members for all four groups are loyal customers. Groups 1, 2, and 4 are also represented by a customer traveling for business purposes and staying in business class, whereas Group 3 is represented by a customer traveling for personal purposes and staying in economy class. In specific,

- Group 1 is represented by a dissatisfied 46-year-old female flying 3,035 miles and experienced a tiny delay on flight departure and arrival; she has average ratings of the airlines' services.
- Group 2 is represented by a satisfied 26-year-old female flying 2,043 miles and departing and arriving on time; she is very satisfied with the airlines' services.
- Group 3 is represented by a dissatisfied 24-year-old male flying 2,501 miles, departing on time, and arriving slightly late; he has mixed ratings of the airlines' services.
- Group 4 is represented by a satisfied 40-year-old female flying 2,602 miles and departing and arriving on time; she is mostly satisfied with the airlines' services.

```
> data[pam_air$medoids, ]
     age flight.distance departure.delay.in.minutes arrival.delay.in.Minutes gender
411   46            3035                         1                        3 Female
2423  26            2043                         0                        0 Female
705   24            2501                         0                        1   Male
1724  40            2602                         0                        0 Female
              cusType      traveltype    class seat depart food gate wifi entertain
411   Loyal Customer Business travel Business    3      3    3    3    3         3
2423  Loyal Customer Business travel Business    5      5    5    5    5         5
705   Loyal Customer Personal Travel      Eco    2      5    2    3    2         2
1724  Loyal Customer Business travel Business    4      4    4    4    4         4
     online booking board legRoom baggage check clean onlineBoard     response
411       3       3     3       3       3     4     3           3 dissatisfied
2423      5       5     5       5       5     4     5           5    satisfied
705       2       2     4       3       5     3     4           2 dissatisfied
1724      4       4     4       4       4     3     4           4    satisfied
```

Figure 8. Medoids of clusters using PAM

Figure 9 shows a graph of the 4 separated clusters. Dots representing individuals in each of the 4 clusters are gathered in their own area with minor overlaps, indicating 4 is the right number for clustering and that our Gower+PAM cluster method is useful.

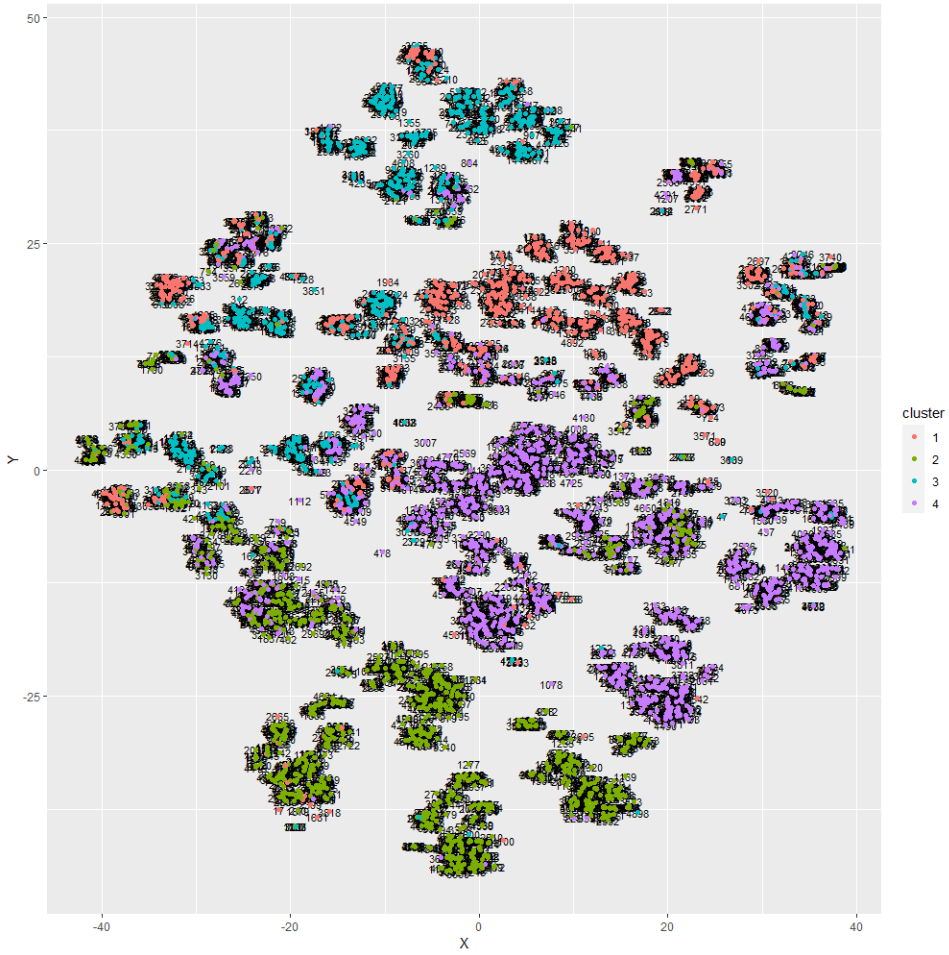

Figure 9. Clusters Separated on Coordinate Axis

MCA+K-Means

Considering the large number of variables we have, in addition to Gower+PAM, an MCA+K-means clustering was carried out to perform dimension reduction and clustering. Using Silhouette width as the criterion, Figure 10 suggests that the best combination was 5 clusters of sizes 1573 (31.5%), 1281 (25.6%), 892 (17.8%), 779 (15.6%), 475 (9.5%) in 4 dimensions, with an average Silhouette width value of 0.15.

```
> best.combination

The best solution was obtained for 5 clusters of sizes 1573 (31.5%), 1281 (25.6%), 892 (17.8%), 779 (15.6%), 475 (9.5%) in 4 dimensions,
or an average Silhouette width value of 0.15.

Cluster quality criterion values across the specified range of clusters (rows) and dimensions (columns):
     X2    X3    X4    X5
3 0.125
4 0.083 0.132
5  0.06 0.128  0.15
6 0.054 0.099 0.147 0.149
7 0.037 0.094 0.123 0.125

The average Silhouette width values of each cluster are:
[1] 0.17 0.20 0.11 0.11 0.15

Cluster centroids:
            Dim.1   Dim.2   Dim.3   Dim.4
Cluster 1  0.0065 -0.0037 -0.0158  0.0026
Cluster 2  0.0170  0.0022  0.0136  0.0021
Cluster 3 -0.0092  0.0000  0.0019 -0.0227
Cluster 4 -0.0177  0.0204  0.0006  0.0101
Cluster 5 -0.0211 -0.0271  0.0111  0.0117

Within cluster sum of squares by cluster:
[1] 0.1526 0.1180 0.1218 0.1043 0.0647
 (between_SS / total_SS =  84.14 %)

Objective criterion value: 13.5227

Available output:

[1] "clusobjbest" "nclusbest"   "ndimbest"    "critbest"    "critgrid"    "crit"
[7] "cluasw"
```

Figure 10. MCA+K-Means Best Combination Output

We also ran a 4-clusters-in-4-dimensions combination and a 5-clusters-in-5-dimensions combination to compare with the best combination. Their outputs are shown below in Figures 11-12.

```
> #--4 clusters & 4 dimensions to compare--#
> mix44=clusmca(data, nclus=4, ndim=4, method=c("clusCA","MCAk"),
+               alphak = .5, nstart = 100, smartStart = NULL, gamma = TRUE,
+               inboot = FALSE, seed = NULL)
  |=============================================================================| 100%> mix44
Solution with 4 clusters of sizes 1707 (34.1%), 1528 (30.6%), 1259 (25.2%), 506 (10.1%) in 4 dimensions.

Cluster centroids:
            Dim.1   Dim.2   Dim.3 Dim.4
Cluster 1  0.0135 -0.0017  0.0125     0
Cluster 2 -0.0078 -0.0153 -0.0081     0
Cluster 3 -0.0181  0.0145  0.0034     0
Cluster 4  0.0230  0.0161 -0.0263     0

Within cluster sum of squares by cluster:
[1] 0.2451 0.1736 0.1311 0.0911
 (between_SS / total_SS =  80.07 %)

Objective criterion value: 10.5849
```

Figure 11. 4 Clusters & 4 Dimensions Output

```
> #--5 clusters & 5 dimensions to compare--#
> mix55=clusmca(data, nclus=5, ndim=5, method=c("clusCA","MCAk"),
+               alphak = .5, nstart = 100, smartStart = NULL, gamma = TRUE,
+               inboot = FALSE, seed = NULL)
  |=============================================================================| 100%
> mix55
Solution with 5 clusters of sizes 1585 (31.7%), 1316 (26.3%), 842 (16.8%), 756 (15.1%), 501 (10%) in 5 dimensions.

Cluster centroids:
            Dim.1   Dim.2   Dim.3   Dim.4 Dim.5
Cluster 1  0.0058  0.0008  0.0174  0.0036     0
Cluster 2  0.0186  0.0001 -0.0140  0.0018     0
Cluster 3 -0.0094  0.0021 -0.0012 -0.0256     0
Cluster 4 -0.0190 -0.0230 -0.0055  0.0088     0
Cluster 5 -0.0227  0.0284 -0.0082  0.0135     0

Within cluster sum of squares by cluster:
[1] 0.2277 0.1975 0.1434 0.1424 0.1003
 (between_SS / total_SS =  80.92 %)

Objective criterion value: 13.4643
```

Figure 12. 5 Clusters & 5 Dimensions Output

External Consistency

The external consistency is checked by Rand Index and an unsupervised tree plot to show the distance among those cluster methods. As shown in Figure 13, all the rand indexes for different pairs are above 0.77, which indicates a high level of similarity between our 4 clustering approaches.

```
> similarity.matrix
          Gower+PAM     MIX54     MIX44     MIX55
Gower+PAM 1.0000000 0.8040899 0.7706188 0.8058508
MIX54     0.8040899 1.0000000 0.8753858 0.9542017
MIX44     0.7706188 0.8753858 1.0000000 0.8672875
MIX55     0.8058508 0.9542017 0.8672875 1.0000000
```

Figure 13. Rand Index Matrix

The unsupervised tree plot in Figure 14 is consistent with the rand index. There are tiny distances among all cluster methods, and the most similar methods are 5-clusters-in-4-dimensions and 5-clusters-in-5-dimensions. Because the Gower+PAM method separates observations into 4 clusters, it is reasonable for it to be farther from the 5-cluster method and closer to the 4-cluster method.

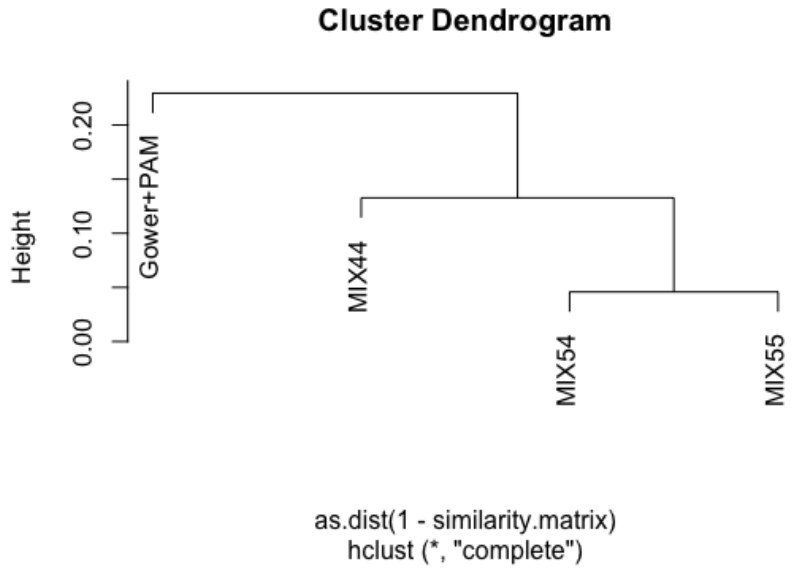

Figure 14. Unsupervised Tree for Clustering Distance Metrics

4. Methodology

The dataset was analyzed using statistical software R. The target variable (Y) is qualitative and binary (satisfied/dissatisfied), and our response variables (X) are mixed variables. Thus, in model formulation, we decided to use binary logistic regression, one pruned tree, bagged tree, random forest, and boosted tree.

- **Binary Logistic regression:** This model is used when the target variable is categorical. The idea of this model is $E(\log(Prob(Y = 1) / Prob(Y = 0))) = \beta 0 + \beta 1x$. In this model, the dependent variable has two possible values and is labeled as "0" and "1"., then model the probability of the certain class. The assumption of the model is no perfect separation and the non-existence of multicollinearity.
- **One pruned tree:** In the classification problem, the tree is used to keep splitting the dataset based on certain values of features that can reduce the entropy most at that stage. The drawback is that this technique may have bias/variance problems.
- **Bagged tree:** This technique is also known as bootstrap aggregation. For bagging, all features are considered for every single split. The classification decision is made through

the majority vote, also called "averaging," which improves predictive accuracy and control over-fitting.

- **Random forest**: This technique is a popular instance of bagging; it grows the deepest tree possible for fitting each subsample. It combines multiple uncorrelated decision trees. The major difference from bagged tree is that the random forest randomly selects certain values of a subset of features for every split. The trees are made independent of one another. The classification decision is also made through the majority vote.
- **Boosted tree:** Each tree in boosting is shallow, not as deep as random forests', and each tree is correlated, and the entire training dataset is used without bootstrapping. It combines several weak learners slowly to achieve a small training error and each weak learner focuses on fixing the mistakes made by the previous weak learner to reduce bias.

5. Modeling

The response variable is whether or not a customer is satisfied with a flight with Invistico Airlines, and the possible predictors include the customer attributes, flight specifics, and the customer's ratings of different services provided by the airlines.

Data cleaning was needed prior to applying analytical techniques. We checked the distribution of service ratings in each customer segment and found a normal distribution. Therefore, we removed all the 0s from the relevant variables as well as the missing values from the column "arrival delay in minutes". To avoid multicollinearity, we also removed the column "departure delay in minutes".

We wanted to maintain balance for the proportions of both classes for training and testing sets. With a random and balanced 75-25 training and test split, we had 3750 observations in the training set and 1250 observations in the test set. Then we set up a 10-fold, 5-run cross-validation set to evaluate the performance of each model we had built and to select the best model.

Logistic Regression

Our first model is a 10-fold cross-validated logistic regression model. According to Figure 15, the cross-validated logistic model has an average accuracy of 0.8969 on different validation sets.

```
> cv.logistic$results
  parameter  Accuracy     Kappa AccuracySD    KappaSD
1      none 0.8968774 0.7918993 0.01436277 0.02892136
```

Figure 15. Performance of logistic regression on validation sets

The parameter estimates and significance given by the logistic model indicate feature importance in the response. We are able to measure the relationships between customer satisfaction and correlated variables and make relevant interpretations. Based on the R output in Figure 16, we select one significant quantitative variable and one significant qualitative variable as examples. The coefficient estimate of arrival delay in minutes is about –0.280. With other

variables held fixed, every one-minute increase in arrival delay reduces the odds of customer satisfaction by about 24% (exp(-0.280)-1). The coefficient estimate of male is about –0.794, and female serves as the reference level for the gender variable. With other variables held fixed, being a male customer reduces the odds of being satisfied by about 55% (exp(-0.794)-1) compared to being a female customer.

```
> summary(cv.logistic$finalModel)

Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8628  -0.2877   0.0230   0.2375   3.5134

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -5.761060   0.556358 -10.355  < 2e-16 ***
air.Age                        -0.084502   0.068068  -1.241 0.214448
air.Flight.Distance            -0.141040   0.071219  -1.980 0.047660 *
air.Arrival.Delay.in.Minutes -0.280349   0.062797  -4.464 8.03e-06 ***
genderMale                     -0.794090   0.130920  -6.065 1.32e-09 ***
cusTypeLoyal Customer           2.999073   0.222653  13.470  < 2e-16 ***
travelTypePersonal Travel      -1.931294   0.200197  -9.647  < 2e-16 ***
classEco                       -0.092891   0.171711  -0.541 0.588528
classEco Plus                  -0.297870   0.256023  -1.163 0.244648
seat2                          -0.880782   0.297325  -2.962 0.003053 **
seat3                          -0.966807   0.304546  -3.175 0.001501 **
seat4                           0.002742   0.282466   0.010 0.992254
seat5                           4.885100   0.636235   7.678 1.61e-14 ***
depart2                        -0.117617   0.340272  -0.346 0.729602
depart3                        -0.855039   0.338969  -2.522 0.011653 *
depart4                        -1.327725   0.305632  -4.344 1.40e-05 ***
depart5                        -2.489961   0.350020  -7.114 1.13e-12 ***
food2                           0.929942   0.356192   2.611 0.009033 **
food3                           1.327441   0.341424   3.888 0.000101 ***
food4                           1.623480   0.331536   4.897 9.74e-07 ***
food5                           1.726620   0.395897   4.361 1.29e-05 ***
gate2                          -0.226842   0.293889  -0.772 0.440195
gate3                           0.053440   0.270793   0.197 0.843557
gate4                          -0.461812   0.269184  -1.716 0.086236 .
gate5                          -0.139387   0.338179  -0.412 0.680214
wifi2                           0.031458   0.308231   0.102 0.918708
wifi3                           0.111766   0.307049   0.364 0.715857
wifi4                           0.258975   0.302177   0.857 0.391427
wifi5                           0.113776   0.315195   0.361 0.718122
entertain2                     -0.594251   0.321962  -1.846 0.064933 .
entertain3                     -0.438982   0.306032  -1.434 0.151449
entertain4                      1.233942   0.286631   4.305 1.67e-05 ***
entertain5                      2.762790   0.343318   8.047 8.46e-16 ***
online2                        -0.103674   0.335118  -0.309 0.757042
online3                        -0.991637   0.311881  -3.180 0.001475 **
online4                        -0.225183   0.307043  -0.733 0.463320
online5                         0.641115   0.327670   1.957 0.050396 .
booking2                        1.668942   0.468104   3.565 0.000363 ***
booking3                        2.187846   0.444005   4.928 8.33e-07 ***
booking4                        2.235117   0.421737   5.300 1.16e-07 ***
booking5                        0.954583   0.456317   2.092 0.036445 *
board2                         -0.119321   0.315834  -0.378 0.705582
board3                          0.449253   0.280118   1.604 0.108758
board4                          0.646897   0.274051   2.361 0.018250 *
board5                          1.275990   0.301148   4.237 2.26e-05 ***
legRoom2                        0.351601   0.298597   1.178 0.238992
legRoom3                        0.252691   0.298383   0.847 0.397069
legRoom4                        1.018249   0.287535   3.541 0.000398 ***
legRoom5                        1.562621   0.300494   5.200 1.99e-07 ***
baggage2                       -0.094246   0.375275  -0.251 0.801708
baggage3                       -0.914353   0.355746  -2.570 0.010163 *
baggage4                        0.076431   0.342373   0.223 0.823349
baggage5                        0.352360   0.358242   0.984 0.325322
check2                          0.461662   0.252350   1.829 0.067331 .
check3                          0.609638   0.219489   2.778 0.005477 **
check4                          0.775505   0.219856   3.527 0.000420 ***
check5                          1.447653   0.259075   5.588 2.30e-08 ***
clean2                         -0.028005   0.395653  -0.071 0.943571
clean3                         -0.674587   0.377193  -1.788 0.073705 .
clean4                         -0.357686   0.367199  -0.974 0.330010
clean5                          0.598521   0.381821   1.568 0.116988
```

Figure 16. Final CV logistic model

Other variables can be interpreted in a similar manner. For instance, Table 3 contains only highly significant variables identified by the logistic model.

Variable	Level	Estimate of Coefficient	P-value
air.Arrival.Delay.in.Minutes	***	-0.280	8.03e-06
genderMale	***	-0.794	1.32e-09
CusTypeLoyal Customer	***	2.999	<2e-16
TraveltypePersonal Travel	***	-1.931	<2e-16
seat5	***	4.885	1.61e-14
depart4	***	-1.328	1.40e-05
depart5	***	-2.490	1.13e-12
food3	***	1.327	0.0001
food4	***	1.623	9.74e-07
food5	***	1.727	1.29e-05
entertain4	***	1.234	1.67e-05
entertain5	***	2.763	8.46e-16
booking2	***	1.669	0.0003
booking3	***	2.188	8.33e-07
booking4	***	2.235	1.16e-07
board5	***	1.276	2.26e-05
legRoom4	***	1.018	0.0004
legRoom5	***	1.563	1.99e-07
check4	***	0.776	0.0004
check5	***	1.448	2.30e-08

Table 3. Highly significant variables, logistic model

From the three plots in Figure 17, we can see that there is no presence of perfect separation between customer satisfaction and each quantitative variable we have. Therefore, we believe the logistic regression model has met the assumption.

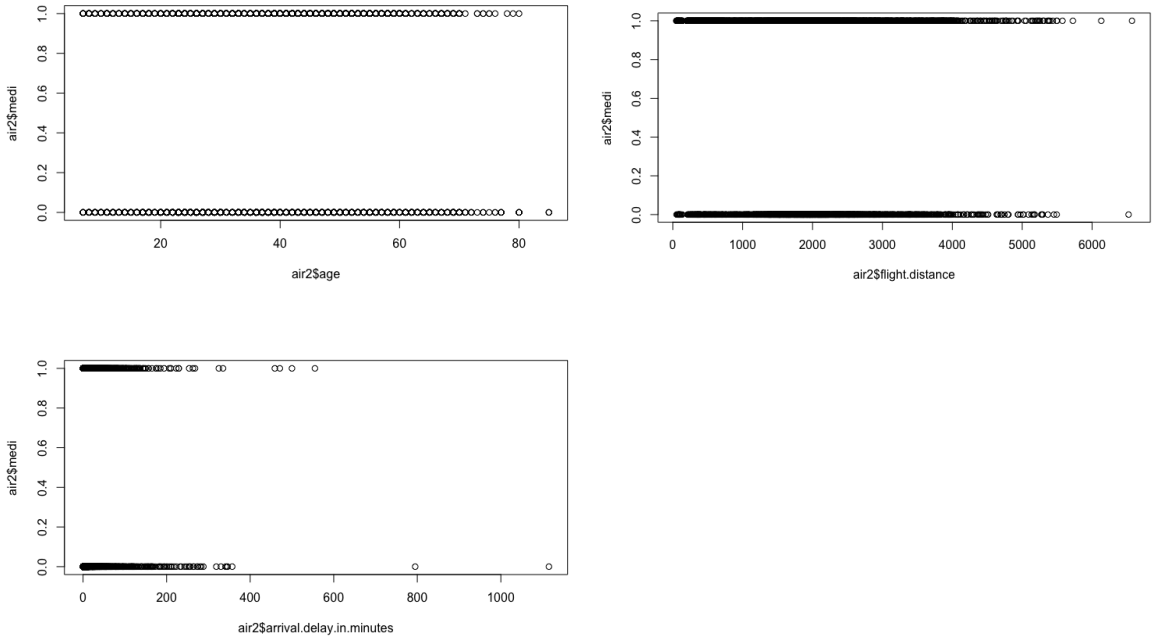

Figure 17. Scatterplots of response vs. age, flight distance, arrival delay

Figure 18 indicates that we needed 8.754 time units elapsed to run the logistic model.

```
> system.time(train(x=trans.tr.pred,y=tr.response,method='glm',trControl=ctrl,family='binomial'))
   user  system elapsed
  8.231   0.495   8.754
```

Figure 18. Time units elapsed, logistic model

One Pruned Tree

Our second model is a 10-fold cross-validated pruned tree. According to Figure 19, the optimal pruned tree is created when the complexity parameter (cp) is about 0.0242. It has the greatest average accuracy of 0.8414 on different validation sets, with a low standard deviation of accuracy.

```
> cv.tree$results
          cp  Accuracy     Kappa AccuracySD    KappaSD
1 0.02423168 0.8414310 0.6770013 0.01930320 0.04195868
2 0.05555556 0.8187158 0.6294613 0.02261268 0.04494377
3 0.58037825 0.6724696 0.2925047 0.12510303 0.29604213
```

Figure 19. Performance of pruned tree on validation sets

Figure 19 indicates that we needed 3.304 time units to run a pruned tree.

```
> system.time(train(x=trans.tr.pred,y=tr.response,method='rpart',trControl=ctrl))
   user  system elapsed
  3.063   0.205   3.304
```

Figure 20. Time units elapsed, pruned tree

According to Figure 21, the pruned tree identifies 5 important variables playing a role in customer satisfaction: entertainment, online booking, online support, seat comfort, and in-flight legroom.

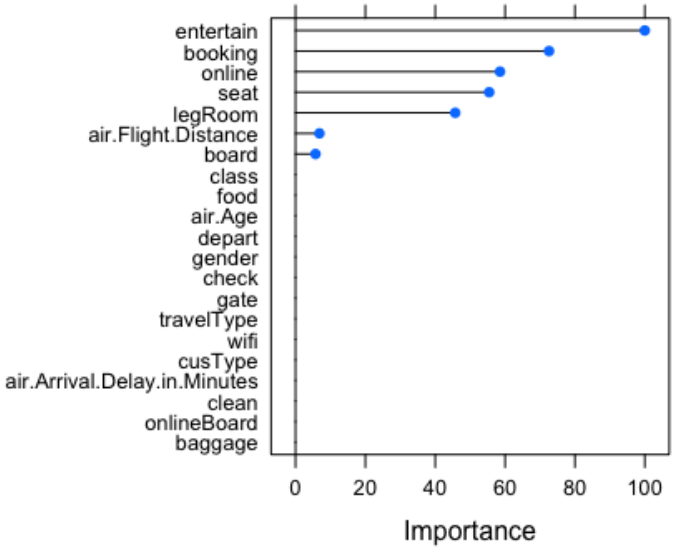

Figure 21. Variable importance diagram, pruned tree

Figure 22 is a plot of the pruned tree. In the top node, the proportion of satisfied customers out of all observations is 55%, so the conditional entropy in that node is about 0.99 (Ent(response)=-0.55*log2(0.55)-0.45*log2(0.45)), which is not a satisfactory number. At the bottom of the tree, 36% of the entire data is in the first terminal node, and about 58% belongs to the third node. Entropy values in those nodes are 0.5 and 0.7, respectively. They are better than the entropy of the top node, which means the tree is good for classification.

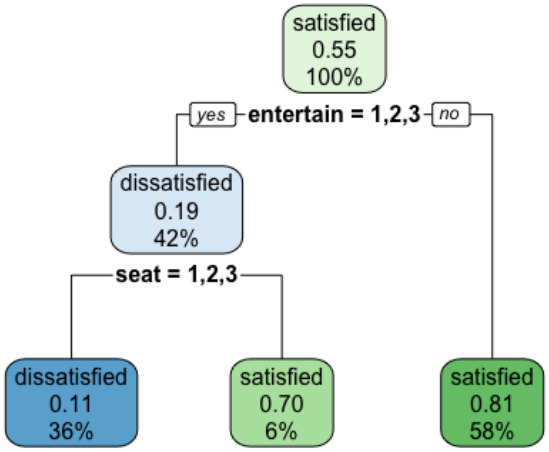

Figure 22. Pruned Tree

Bagged Tree

We then used the bootstrapped aggregation method to build a bagged tree. According to Figure 23, it has an average validation accuracy of 0.9140, with a low standard deviation of accuracy.

```
> cv.bagged.tree$results
  parameter  Accuracy     Kappa AccuracySD    KappaSD
1      none 0.9139968 0.8264249 0.01258137 0.02529176
```

Figure 23. Performance of bagged tree on validation sets

Figure 24 indicates that we needed 29.762 time units to run a bagged tree.

```
> system.time(train(x=trans.tr.pred,y=tr.response,method='treebag',trControl=ctrl))
user  system elapsed
27.789   1.763  29.762
```

Figure 24. Time units elapsed, bagged tree

According to Figure 25, bagged tree identifies 5 important variables playing a role in customer satisfaction: entertainment, online booking, online support, seat comfort, and in-flight legroom.

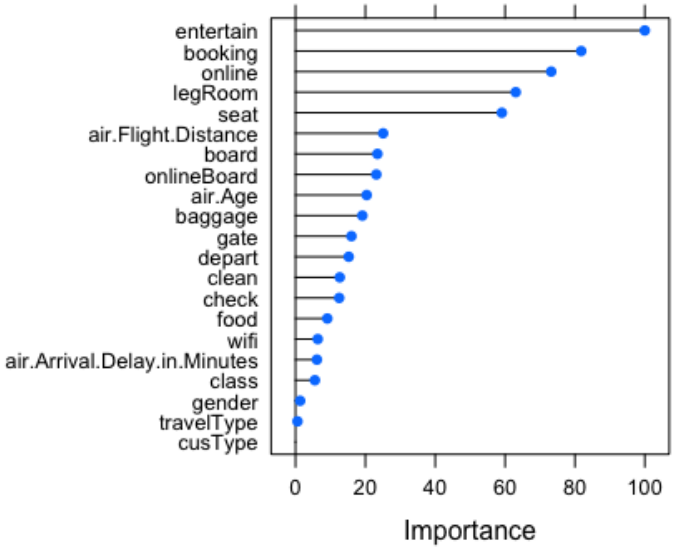

Figure 25. Variable importance diagram, bagged tree

Random Forest

Figure 26 suggests that the ideal number of variables available for splitting at each tree node is 11. This leads to the greatest average accuracy on different validation sets, which is 0.9201, and a low standard deviation of accuracy.

```
> cv.randomforest$results
  mtry  Accuracy     Kappa AccuracySD    KappaSD
1    2 0.9094691 0.8168378 0.01521002 0.03087390
2   11 0.9201307 0.8386270 0.01373071 0.02771723
3   21 0.9172507 0.8328896 0.01383868 0.02788823
```

Figure 26. Performance of random forest on validation sets

Figure 27 indicates that we needed 5134.038 time units to run the random forest model.

```
> system.time(train(x=trans.tr.pred,y=tr.response,method='cforest',trControl=ctrl))
 user   system  elapsed
5072.942   26.567 5134.038
```

Figure 27. Time units elapsed, random forest

According to Figure 28, random forest identifies 2 important variables playing a role in customer satisfaction: entertainment and seat comfort.

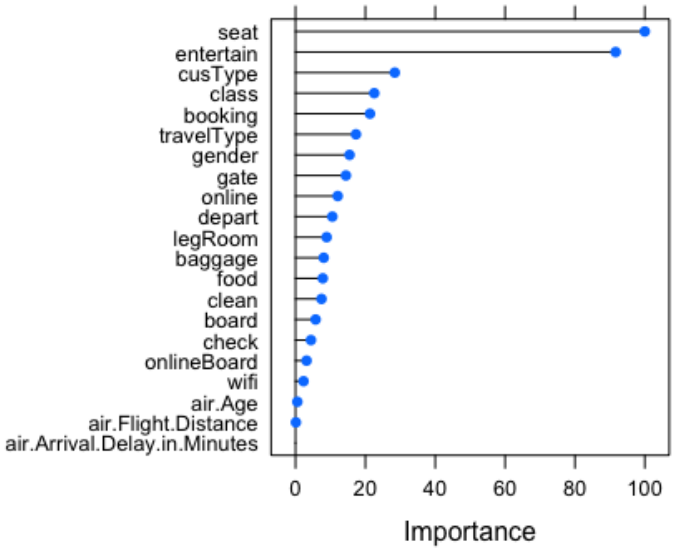

Figure 28. Variable importance diagram, random forest

Boosted Tree

Figure 29 below demonstrates the performance of boosted trees in the process of cross-validation; the trees produced are less complicated and correlated with one another. Given that we do not have sample replicates and cannot use them to fit trees like what we did for the bagged tree or random forest (bagging method), we have more parameters here, including shrinkage,

depth of the tree, n.minobsinnode, and the number of trees used. Among all the choices below, the bagged tree with shrinkage=0.1, interaction.depth=3, n.minobsinnode=10, n.trees=150 is the best. The average validation accuracy is 0.9165, with a low standard deviation of accuracy.

```
> cv.boosted$results
  shrinkage interaction.depth n.minobsinnode n.trees  Accuracy     Kappa AccuracySD    KappaSD
1       0.1                 1             10      50 0.8643595 0.7263684 0.01798427 0.03635662
4       0.1                 2             10      50 0.8872340 0.7722501 0.01623441 0.03272956
7       0.1                 3             10      50 0.8973105 0.7926425 0.01502026 0.03036598
2       0.1                 1             10     100 0.8841431 0.7662008 0.01722136 0.03458410
5       0.1                 2             10     100 0.9027502 0.8036413 0.01505240 0.03037735
8       0.1                 3             10     100 0.9109600 0.8202396 0.01353719 0.02724395
3       0.1                 1             10     150 0.8915555 0.7811577 0.01497481 0.03006305
6       0.1                 2             10     150 0.9080274 0.8143357 0.01452218 0.02923613
9       0.1                 3             10     150 0.9164509 0.8314494 0.01460163 0.02939447
```

Figure 29. Performance of boosted tree on validation sets

Figure 30 indicates that we needed 54.098 time units to run the boosted tree.

```
>system.time(train(x=trans.tr.pred,y=tr.response,method='gbm',trControl=ctrl))
user  system elapsed
53.453   0.965  54.098
```

Figure 30. Time units elapsed, boosted tree

According to Figure 31, the boosted tree identifies entertainment as an important variable in influencing customer satisfaction.

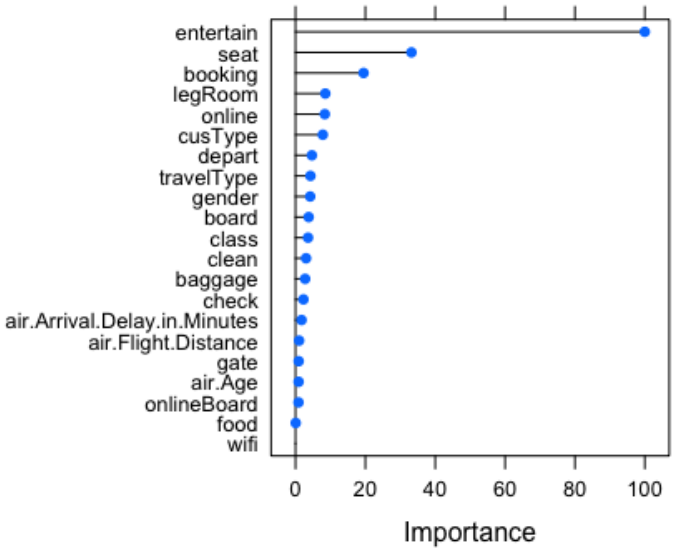

Figure 31. Variable importance diagram, boosted tree

As opposed to logistic regression, trees and their variations provide neither coefficient estimates nor significance. As a result, we are unable to measure how one attribute influences a customer's satisfaction level specifically. But classification trees are easy to visualize and interpret.

Comparing Models

Table 4 compares the results of different cross-validated models. While random forest has the highest accuracy on the validation sets, we also need to take into account the time spent on building each model. Random forest has relatively weak practical significance due to the large amount of time that it took to run. The boosted tree and bagged tree have similarly high accuracy values. Nevertheless, the time elapsed on the bagged tree is nearly half that on the boosted tree. Therefore, both models were considered the best choices and tested on the test set.

Model	Validation Accuracy	Time Elapsed
Best Random Forest	0.9201307	5134.038
Best Boosted Tree	0.9164509	54.098
Bagged Tree	0.9139968	29.762
Logistic Regression	0.8968774	8.754
Best One Pruned Tree	0.8414310	3.304

Table 4. Comparison of five cross-validated models

Figure 32 shows how well the bagged tree classifies observations in the test set. It has a 0.9119 test accuracy with a tight 95% confidence interval. That is to say, the test error is 1-0.9119 = 8.81%. On the other hand, a 0.5492 No information rate (NIR) indicates that the classes of the observations are nearly evenly distributed. Tiny p-value [acc>NIR] rejects the null hypothesis that accuracy is not better than NIR. Also, we are able to correctly identify 91.25% satisfied customers as satisfied (sensitivity) and 91.12% dissatisfied customers as dissatisfied (specificity).

```
> confusionMatrix(data=pred.bagged.tree,test.response,positive = "satisfied")
Confusion Matrix and Statistics

          Reference
Prediction     dissatisfied satisfied
  dissatisfied          513        60
  satisfied              50       626

               Accuracy : 0.9119
                 95% CI : (0.8948, 0.9271)
    No Information Rate : 0.5492
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8224

 Mcnemar's Test P-Value : 0.3908

            Sensitivity : 0.9125
            Specificity : 0.9112
         Pos Pred Value : 0.9260
         Neg Pred Value : 0.8953
             Prevalence : 0.5492
         Detection Rate : 0.5012
   Detection Prevalence : 0.5412
      Balanced Accuracy : 0.9119

       'Positive' Class : satisfied
```

Figure 32. Performance of CV bagged tree on test set

Similarly, Figure 33 shows the boosted tree's performance on the test set. It has a 0.9239 test accuracy with a tight 95% confidence interval. That is to say, the test error is 1-0.9239 = 7.61%. Also, we are able to correctly identify 93.29% satisfied customers as satisfied (sensitivity) and 91.3% dissatisfied customers as dissatisfied (specificity). The boosted tree has better accuracy measures on the test set than the bagged tree does.

```
> confusionMatrix(data=pred.boost,test.response,positive = "satisfied")
Confusion Matrix and Statistics

          Reference
Prediction     dissatisfied satisfied
  dissatisfied          514        46
  satisfied              49       640

               Accuracy : 0.9239
                 95% CI : (0.9078, 0.938)
    No Information Rate : 0.5492
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8463

 Mcnemar's Test P-Value : 0.8374

            Sensitivity : 0.9329
            Specificity : 0.9130
         Pos Pred Value : 0.9289
         Neg Pred Value : 0.9179
             Prevalence : 0.5492
         Detection Rate : 0.5124
   Detection Prevalence : 0.5516
      Balanced Accuracy : 0.9230

       'Positive' Class : satisfied
```

Figure 33. Performance of CV boosted tree on test set

6. Conclusions

Customer satisfaction is critical in revenue generation and brand reputation across all business-related industries, especially the airline industy. With a focus on helping Invistico Airlines improve their overall customer satisfaction level, this study first analyzed a random sample of 5000 customers' demographics and their ratings on airline services by clustering using Gower+PAM and MCA+K-means methods, then we developed predictive models to estimate potential customer's satisfaction status with regression model and classification trees. The choice of the best model depends on the expectations of Invistico Airlines. If the primary focus is higher efficiency, we would recommend the bagged model. Otherwise, we would go for the boosted model, which takes a longer time to run but achieves a satisfying performance out of all models available.

Moreover, we wanted to investigate what aspects should be emphasized to improve customer satisfaction. Based on the variable importance diagrams derived from different tree models, in-flight entertainment turns out to be the most essential factor in improving customer satisfaction. Therefore, we would recommend that Invistico Airlines focus on this service to improve customer satisfaction. For example, they can provide more movies and TV programs.

Seat comfort is another service that Invistico Airlines should pay attention to for the realization of more satisfying customer experiences. Therefore, we would recommend the airlines improve these two services, especially entertainment programs because they are easier and more cost-efficient to implement than improving seats on airplanes. Meanwhile, the logistic model identified some variables, such as customer's gender and travel type, at the 5% significance level. But some were not considered important by the tree related models. Further research on different combinations of customer attributes may extend the findings regarding whether or not certain customer segments should be targeted by the airlines.

In this analysis, we did not consider H-measure or severity ratio, which are relevant to the total cost that the company needs to pay. Instead, we chose the best model based on cross-validation accuracy and time elapsed. It is also critical to find the most cost-efficient model by observing how H-measure changes with different severity ratios. A key extension of our work is to also focus on minimizing misclassification errors through comparing metrics in the H-measure result table as well as the ROC plot. In the context of the problem, false positives refer to misclassifying a dissatisfied customer as satisfied and may impair customer experiences. False negatives, on the other hand, refer to misclassifying a satisfied customer as dissatisfied and can cause loss of business opportunities. The most appropriate severity ratio is subject to the business objectives of the company. Comparing both H-measure and accuracy metrics allows us to see if the choice of the best model will keep consistent. Finally, to help the airlines decide how to lower the proportion of dissatisfied customers, we can use partial dependency plots to see how each important variable influences the response variable, and we can also use interaction plots to see how the important variables interact with non-important variables or other important variables.

References

Effler, G. (2021, May 12). 2021 North America Airline Satisfaction Study. J.D. Power. Retrieved December 12, 2021, from https://www.jdpower.com/business/press-releases/2021-north-america-airline-satisfaction-study

Gourdin, K. N. (2013). The Evolving Relationship Between Airline Profitability and Passenger Satisfaction. Journal of Transportation Management, 24(1), 7+. https://link.gale.com/apps/doc/A634040083/AONE?u=mlin_oweb&sid=googleScholar&xid=96d9a96b

Airlines Customer satisfaction. (2020, March 19). Kaggle. https://www.kaggle.com/sjleshrac/airlines-customer-satisfaction