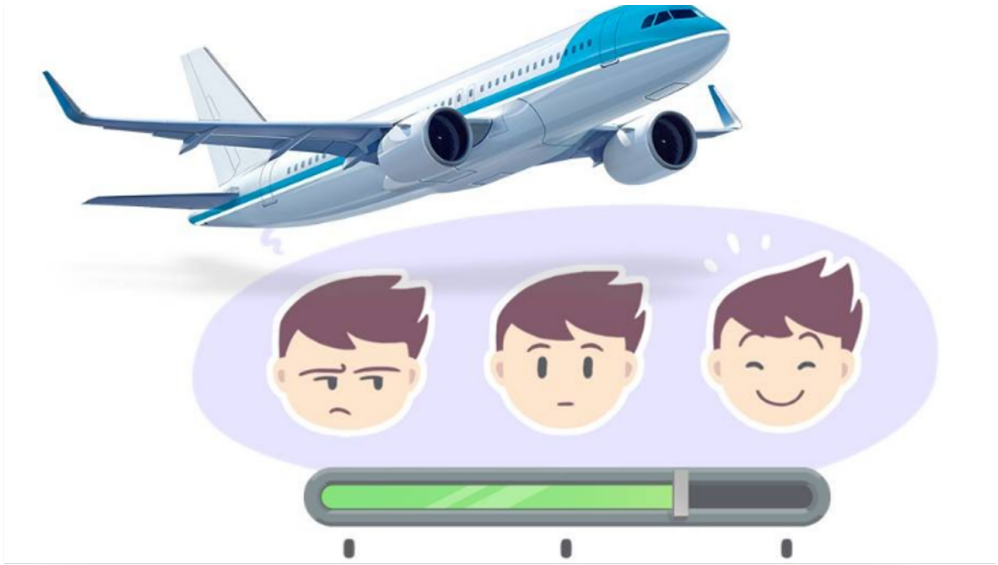
Airline Customer Satisfaction Analysis and Prediction



Contents

Introduction

Traveling by plane has become more popular nowadays due to its comfort, convenience and short travel time. However, different customers traveling on different flights with different experiences could result in different satisfaction levels.

Our research topic for this project is "Airline Customer Satisfaction Analysis and Prediction." This project aims to build customer satisfaction prediction models and pick the one with the best performance to help airline companies achieve business success. We will try to build different types of models and compare the results, then try to optimize the models by applying the techniques we learned in class. We aim to use the 23 variables in the data set to predict passenger satisfaction and suggest to airlines whether a future customer would be satisfied with their service given the details of the other parameter values.

Dataset

The data set is open source from Kaggle, originally collected initially by an airline organization[1]. The Data set contains close to 129,881 sampling data points of the details of customers who have already flown with them. The customers' feedback on various contexts and flight data has been consolidated.

Qualitative variables in the collected dataset:

- Satisfaction: Airline satisfaction level (Satisfaction, dissatisfaction) - Response variable.
- Gender: Gender of the passengers (Female, Male).
- Customer Type: The customer type (Loyal customer, disloyal customer).
- Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel).
- Class: Travel class in the plane the passengers (Business, Eco, Eco Plus).

Quantitative variables in the collected dataset:

- Flight distance: The flight distance of this journey.
- Age: The actual age of the passengers.
- Departure Delay in Minutes: Minutes are delayed when departure.
- Arrival Delay in Minutes: Minutes delayed when the Arrival.
- In flight Wi-Fi service: Satisfaction level of the inflight Wi-Fi service.
- Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient.
- Ease of Online booking: Satisfaction level of online booking.
- Gate location: Satisfaction level of Gate location.
- Food and drink: Satisfaction level of Food and drink.
- Online boarding: Satisfaction level of online boarding.
- Seat comfort: Satisfaction level of Seat comfort.
- Inflight entertainment: Satisfaction level of inflight entertainment.
- On-board service: Satisfaction level of Onboard service.

- Legroom service: Satisfaction level of Leg room service.
- Baggage handling: Satisfaction level of baggage handling.
- Check-in service: Satisfaction level of Check-in service.
- Inflight service: Satisfaction level of inflight service.
- Cleanliness: Satisfaction level of Cleanliness.

Note: The Satisfaction level goes from 0-5, where '0' stands for "Not Applicable," '1' is "Least Satisfied," and '5' is "Most Satisfied."

License concerns for the data set: Open data set from "Kaggle", collected initially by an airline organization. The actual name of the company is not given due to various purposes that's why the name is Invistico airlines.

Data Wrangling

Before we start our project, We load all the packages that we will need for this project here:

First of all, let us take a look at the data set by reading in the csv file:

```
##   satisfaction Gender  Customer.Type Age  Type.of.Travel    Class
## 1    satisfied Female Loyal Customer  65 Personal Travel      Eco
## 2    satisfied   Male Loyal Customer  47 Personal Travel Business
## 3    satisfied Female Loyal Customer  15 Personal Travel      Eco
##   Flight.Distance Seat.comfort Departure.Arrival.time.convenient Food.and.drink
## 1             265            0                                 0              0
## 2            2464            0                                 0              0
## 3            2138            0                                 0              0
##   Gate.location Inflight.wifi.service Inflight.entertainment Online.support
## 1             2                     2                      4              2
## 2             3                     0                      2              2
## 3             3                     2                      0              2
##   Ease.of.Online.booking On.board.service Leg.room.service Baggage.handling
## 1                      3                3                0                 3
## 2                      3                4                4                 4
## 3                      2                3                3                 4
##   Checkin.service Cleanliness Online.boarding Departure.Delay.in.Minutes
## 1               5           3               2                         0
## 2               2           3               2                       310
## 3               4           4               2                         0
##   Arrival.Delay.in.Minutes
## 1                        0
## 2                      305
## 3                        0
```

```
##  [1] "satisfaction"                     "Gender"
##  [3] "Customer.Type"                    "Age"
##  [5] "Type.of.Travel"                   "Class"
##  [7] "Flight.Distance"                  "Seat.comfort"
##  [9] "Departure.Arrival.time.convenient" "Food.and.drink"
## [11] "Gate.location"                    "Inflight.wifi.service"
## [13] "Inflight.entertainment"           "Online.support"
```

```
## [15] "Ease.of.Online.booking"            "On.board.service"
## [17] "Leg.room.service"                  "Baggage.handling"
## [19] "Checkin.service"                   "Cleanliness"
## [21] "Online.boarding"                   "Departure.Delay.in.Minutes"
## [23] "Arrival.Delay.in.Minutes"
```

Since some of the columns that contains the ratings for the airline company in various areas of services, the rating goes from 0-5, where '0' stands for "Not Applicable," '1' is "Least Satisfied," and '5' is "Most Satisfied". Since "0" stands for "Not Applicable", doesn't necessary mean the service was bad, so in order to prevent miss-leading for our model later, we want to convert those "0" rating to be "NA", then drop those rows that contains "NA". See below R chunks for the data cleaning.

```
##                                    na_count
## satisfaction                              0
## Gender                                    0
## Customer.Type                             0
## Age                                       0
## Type.of.Travel                            0
## Class                                     0
## Flight.Distance                           0
## Seat.comfort                           4797
## Departure.Arrival.time.convenient      6664
## Food.and.drink                         5945
## Gate.location                             2
## Inflight.wifi.service                   132
## Inflight.entertainment                 2978
## Online.support                            1
## Ease.of.Online.booking                   18
## On.board.service                          5
## Leg.room.service                        444
## Baggage.handling                          0
## Checkin.service                           1
## Cleanliness                               5
## Online.boarding                          14
## Departure.Delay.in.Minutes                0
## Arrival.Delay.in.Minutes                393
```

Check out how many rows we have left in our data set after dropping all the "NA" values:

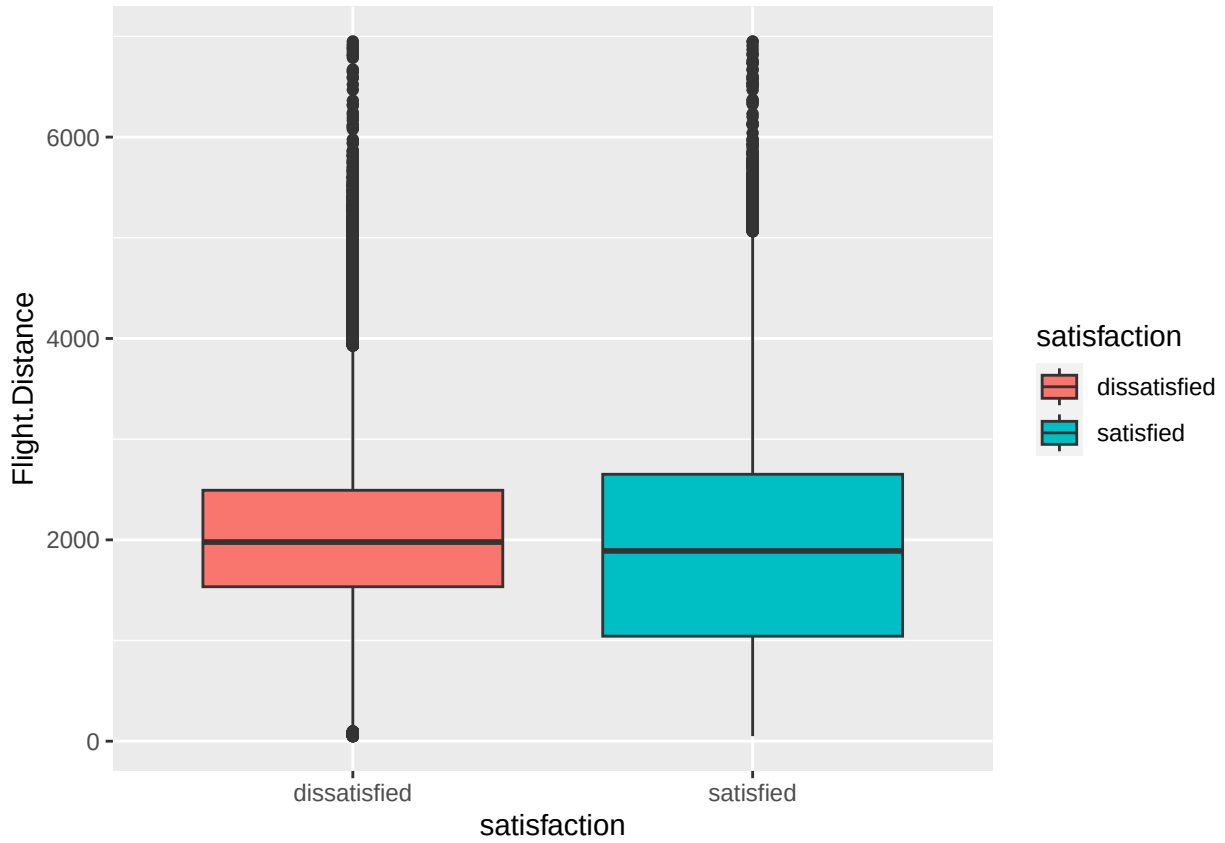
```
## [1] 119255
```

Data Exploration

In order to build better model, first of all, we should get familiar with our data set and get a better understanding of the data context.

Flight.Distance vs Satisfaction

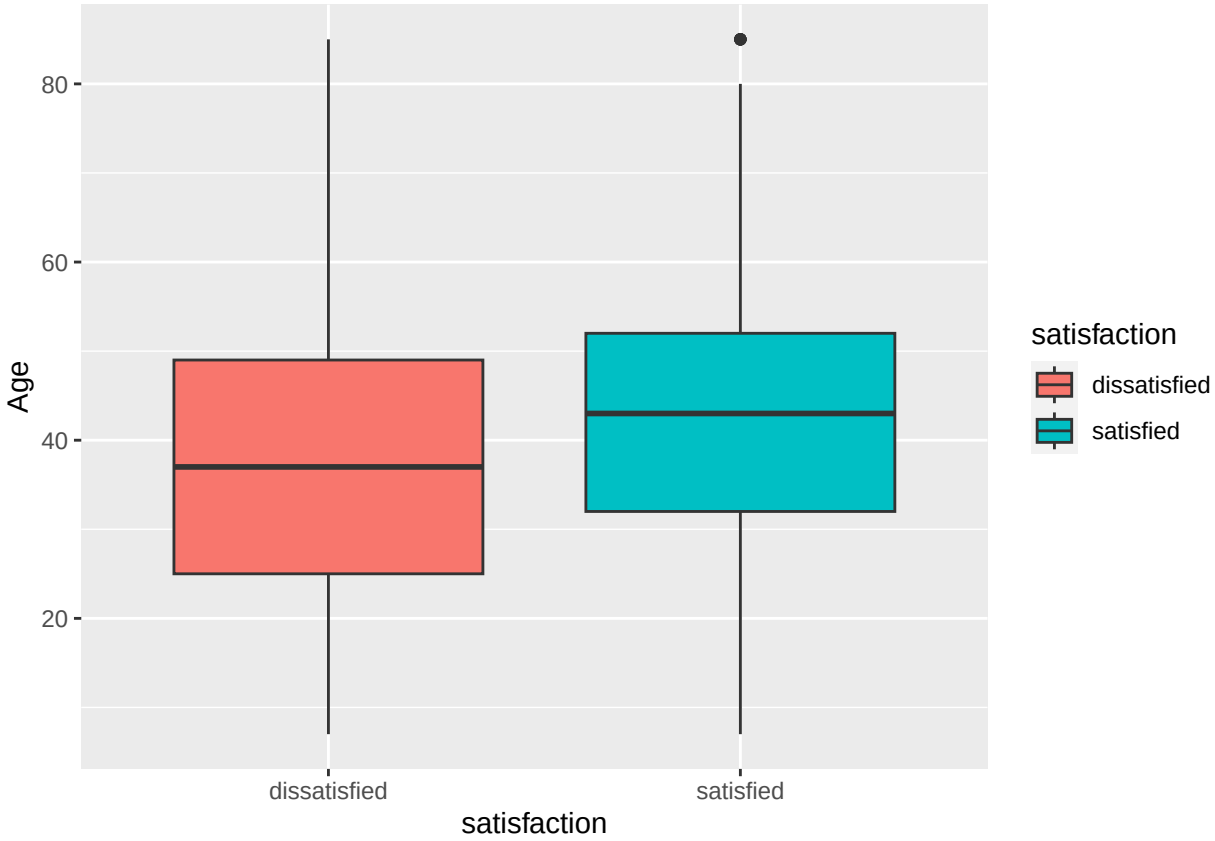
Plot Flight.Distance vs satisfaction box plot to see if the Flight Distance would affect the satisfaction:



As we can see from the box plot above, the median "Flight Distance" for dissatisfied category and satisfied category are almost at the same level, which means "Flight Distance" does not vary significantly based on the satisfaction category.

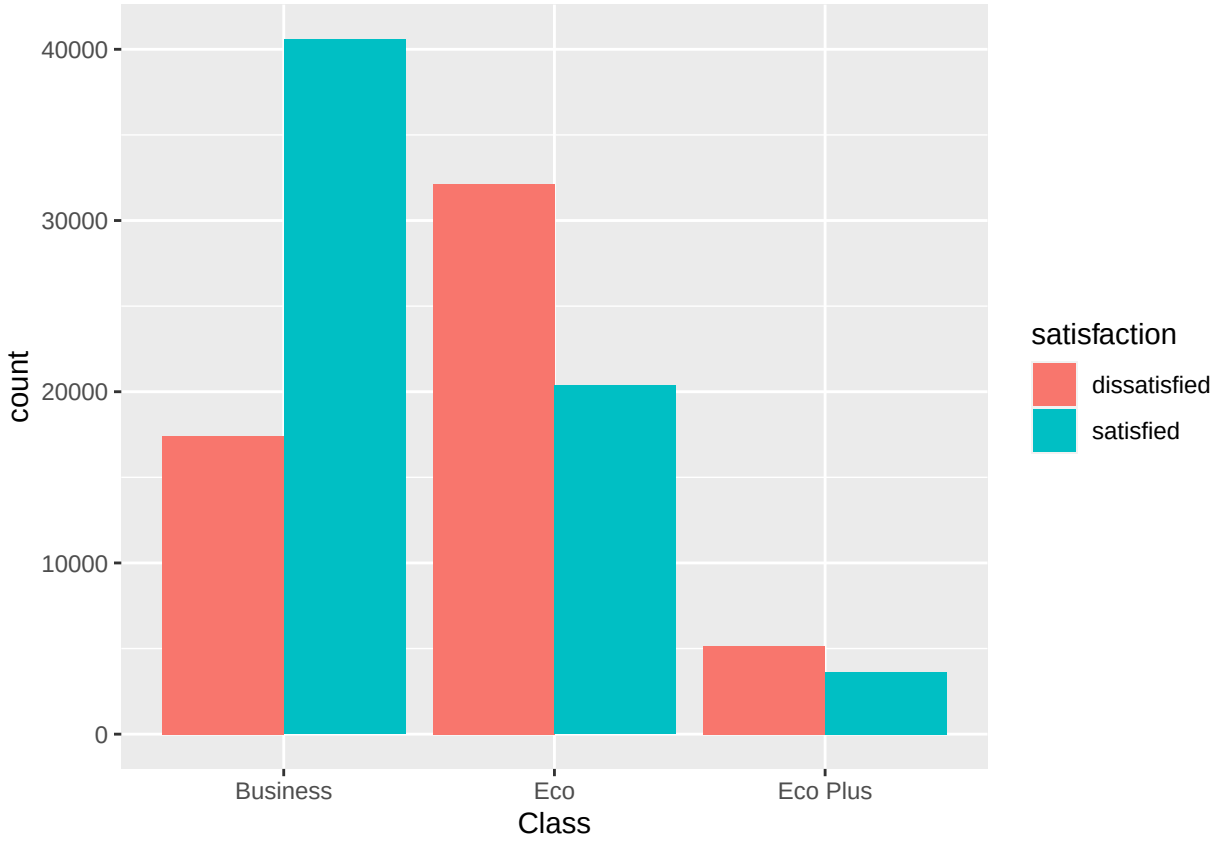
Age vs Satisfaction

Plot Age vs satisfaction box plot to see if the Age would affect the satisfaction:



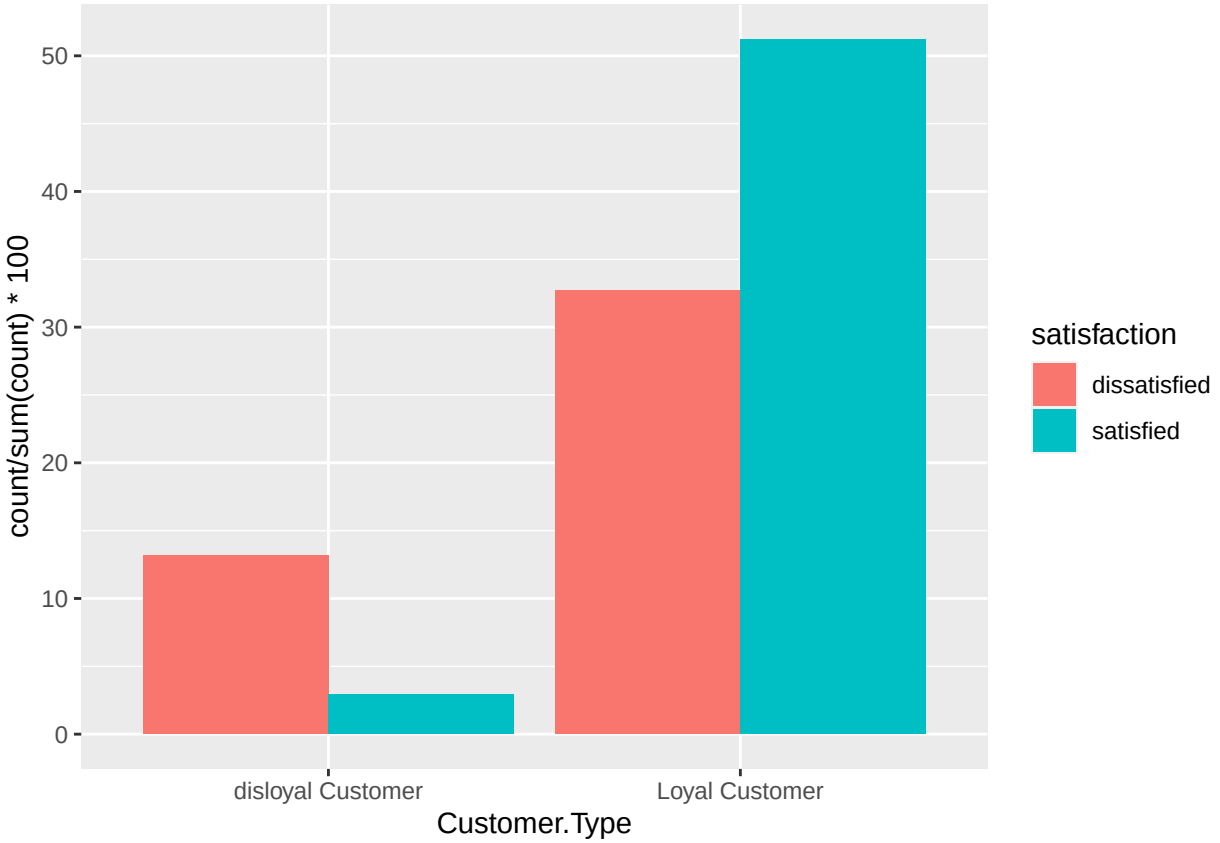
From the box plot above, we could observe that satisfied customer tend to have older age than the dissatisfied customer.

Plot the bar plot based on the type of “Class”:



Based on the bar plot above, we could clearly see that the flight class type has significant impact on customer satisfaction. The "Business" class tends to have the highest satisfaction ratio, while the "Eco" class tends to have the lowest satisfaction ratio.

Plot the bar plot based on the type of "Customer Loyalty":



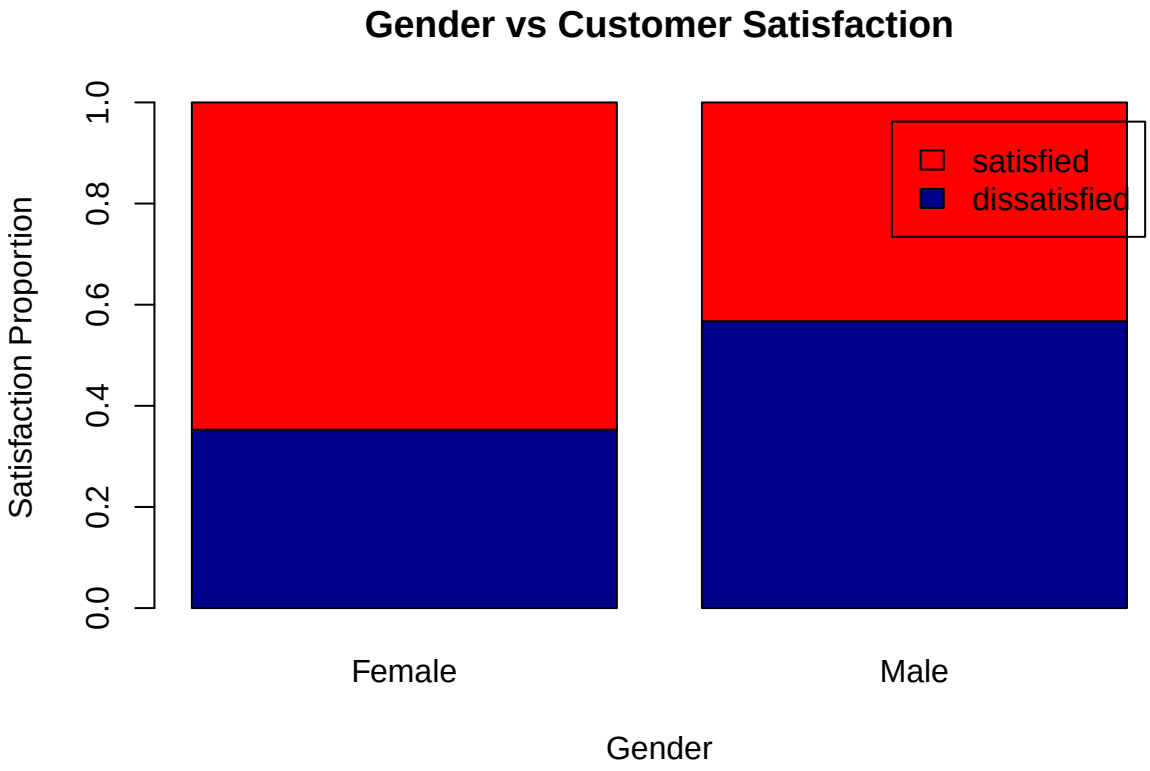
Based on the bar plot above, we could clearly see that the customer loyalty type also has significant impact on customer satisfaction. The loyal customers tend to have much higher satisfaction ratio than the disloyal customers.

Data Sampling

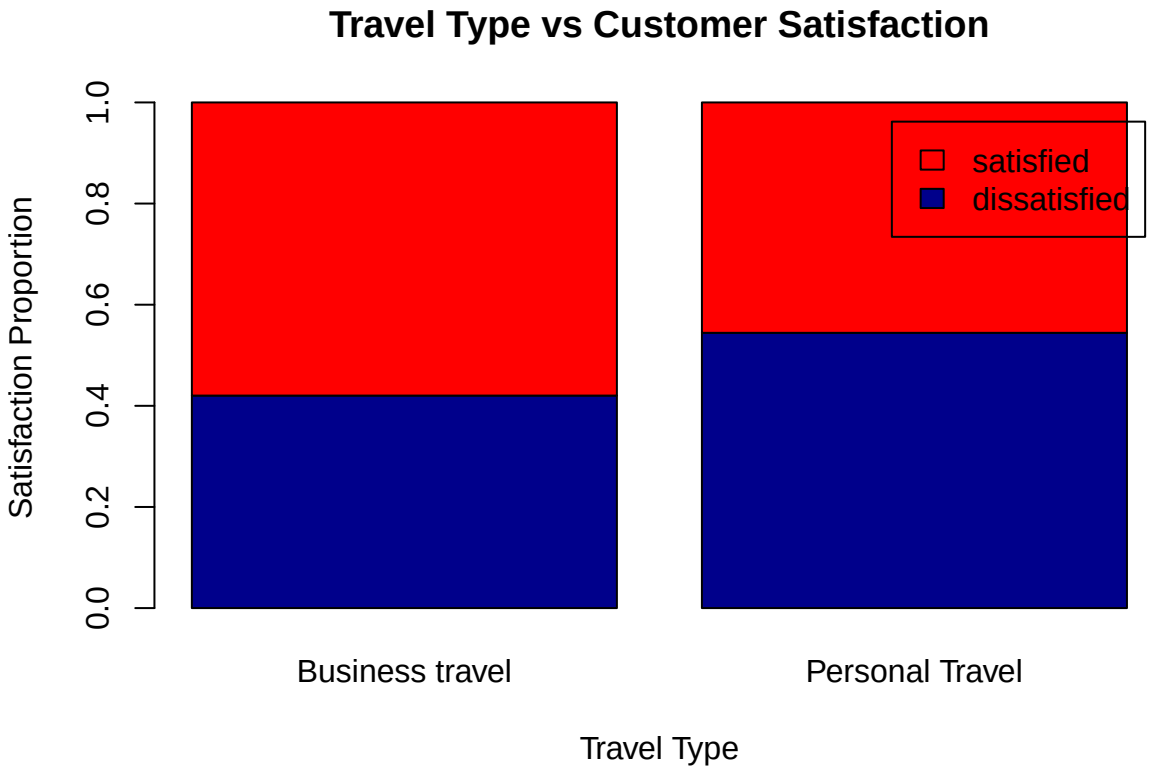
Select variable that should be used as "Strata" for stratified sampling

We want to apply the stratified sampling technique we learned during the course. First of all, we want to check how the satisfied and dissatisfied distribution proportions are based on each categorical variable, see if we could find a categorical variable as our strata for sampling:

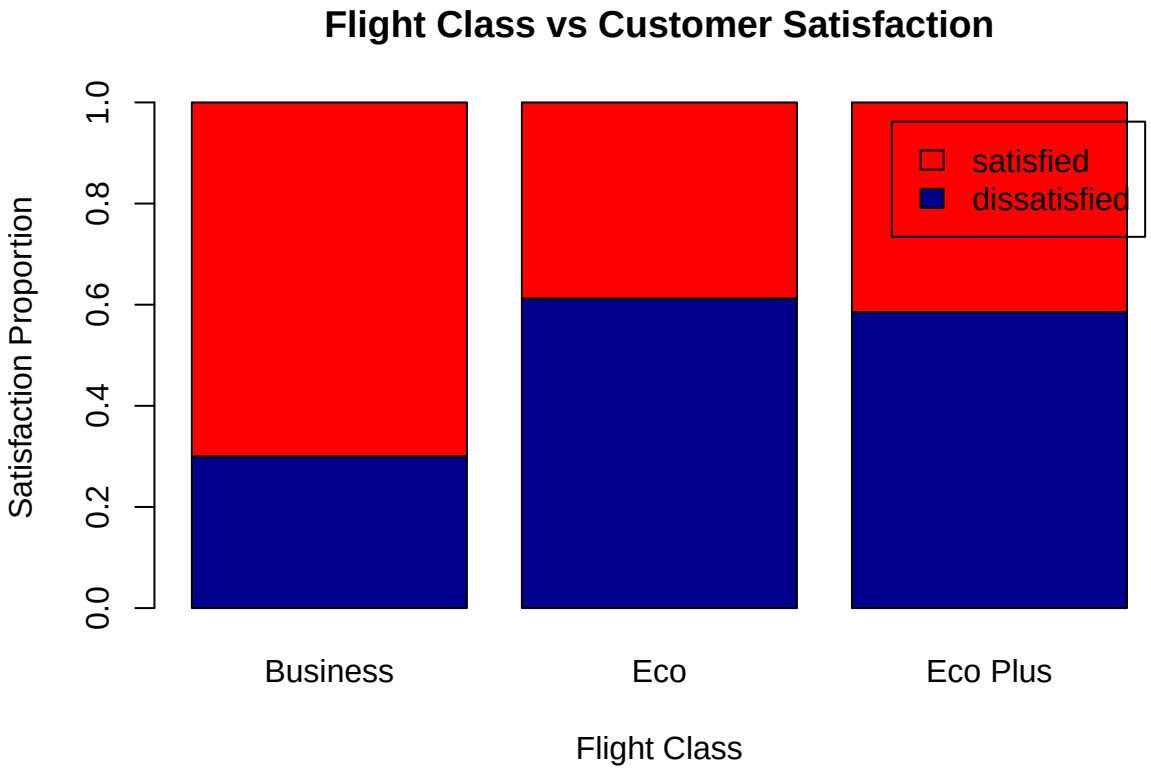
Let us take a look at the proportion of the satisfaction proportion in each Gender:



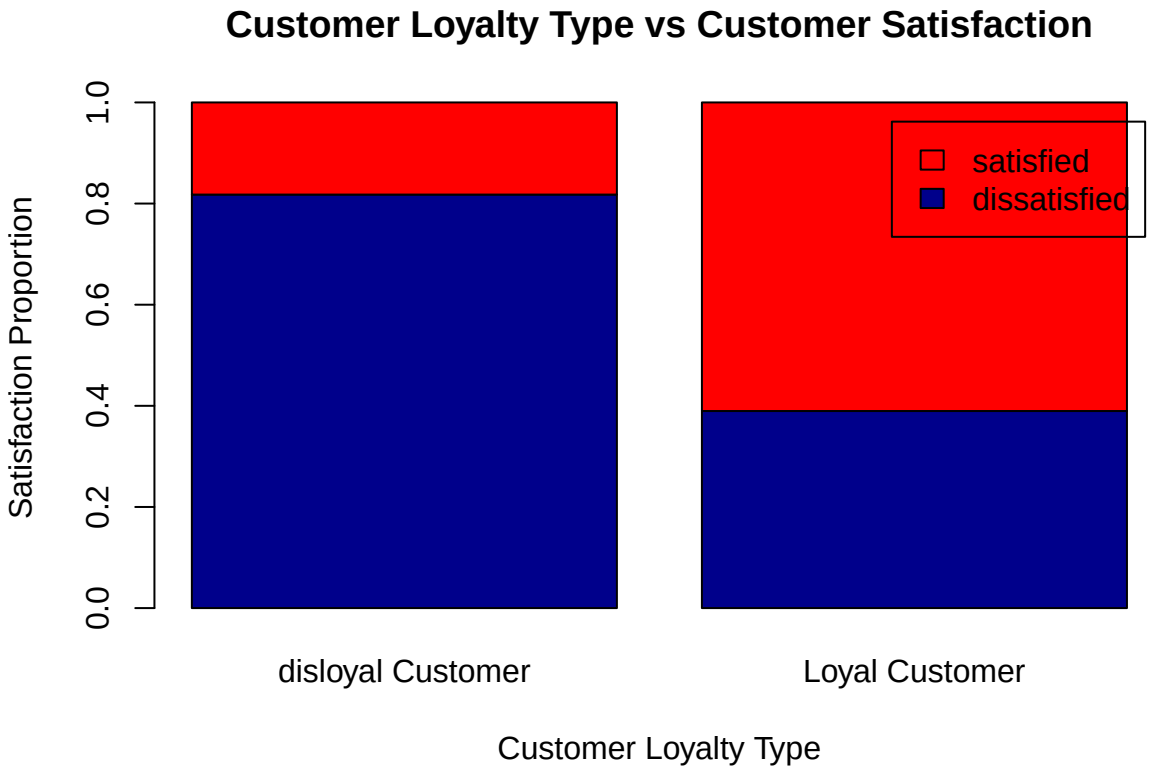
Let us take a look at the proportion of the satisfaction proportion in each of the travel type:



Let us take a look at the proportion of the satisfaction proportion in each of the flight class type:



Let us take a look at the proportion of the satisfaction proportion in each of the customer loyalty type:



We could observe clear customer satisfaction ratio varies significantly when "Customer Loyalty Type" changes. Therefore, we decided to apply the "Stratified Sampling Technique" to sample the units from our data set with "Customer Loyalty Type" as the strata.

First of all, we should check out how many units are in each "Customer Loyalty Type":

```
##
## disloyal Customer    Loyal Customer
##             19180            100075
```

We use the "contrasts" method to check the class categories in the "satisfaction" column/variable:

```
##              satisfied
## dissatisfied         0
## satisfied            1
```

Create Train/Test Data and Folds used for Cross-validation

We create the "train"(80%) and "test"(20%) data set for training and testing the models we will be building in the project, during this sampling process for train and test data, we decided to apply the "Stratified Sampling Technique" to sample the units with airline "Customer Loyalty Type" as the strata:

```
##
## disloyal Customer    Loyal Customer
##             15344             80060
```

```
##
## disloyal Customer    Loyal Customer
##              3836             20015
```

We also created the 10 folds from our data set using the "stratified k-fold cross validation" technique, and we are planning to apply the same K folds throughout all the models.

Methodology

Logistic Regression

Since the response variable was a binary classification nature, logistic regression was chosen as the first algorithmic method to be used. We have started with the full model with all the predictor variables on the training dataset.

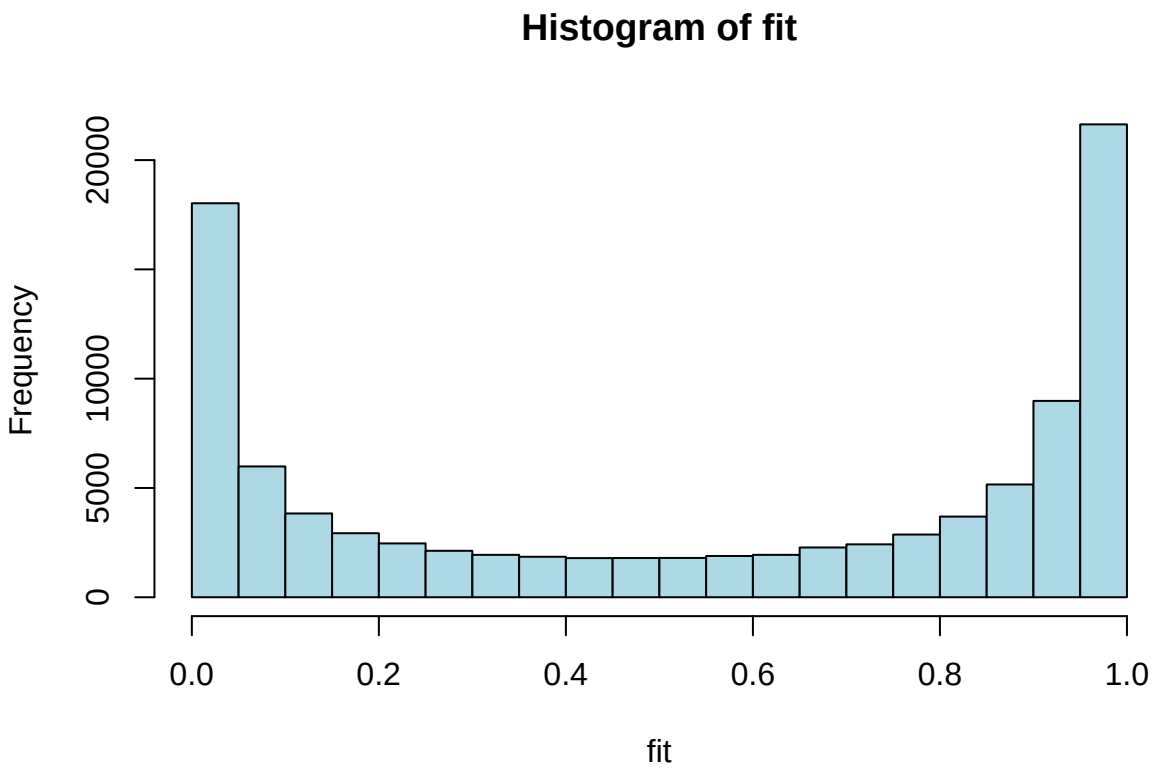
```
##
## Call:
## glm(formula = satisfaction ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3169  -0.3738   0.1308   0.4204   3.7067
##
## Coefficients:
##                                    Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                      -1.011e+01  9.601e-02 -105.289  < 2e-16 ***
## GenderMale                       -9.497e-01  2.168e-02  -43.815  < 2e-16 ***
## Age                              -3.504e-03  7.490e-04   -4.679 2.89e-06 ***
## Type.of.TravelPersonal Travel    -1.033e+00  3.087e-02  -33.460  < 2e-16 ***
## ClassEco                         -7.261e-01  2.838e-02  -25.582  < 2e-16 ***
## ClassEco Plus                    -9.100e-01  4.286e-02  -21.232  < 2e-16 ***
## Flight.Distance                  -9.064e-05  1.107e-05   -8.189 2.63e-16 ***
## Seat.comfort                      5.928e-01  1.267e-02   46.802  < 2e-16 ***
## Departure.Arrival.time.convenient -4.391e-01  1.168e-02  -37.604  < 2e-16 ***
## Food.and.drink                    1.910e-01  1.345e-02   14.201  < 2e-16 ***
## Gate.location                    -5.679e-02  1.136e-02   -4.999 5.77e-07 ***
## Inflight.wifi.service            -9.305e-02  1.132e-02   -8.223  < 2e-16 ***
## Inflight.entertainment            8.866e-01  1.165e-02   76.110  < 2e-16 ***
## Online.support                    9.360e-02  1.193e-02    7.848 4.23e-15 ***
## Ease.of.Online.booking            2.891e-01  1.523e-02   18.973  < 2e-16 ***
## On.board.service                  3.763e-01  1.111e-02   33.864  < 2e-16 ***
## Leg.room.service                  2.926e-01  9.584e-03   30.526  < 2e-16 ***
## Baggage.handling                  1.368e-01  1.255e-02   10.903  < 2e-16 ***
## Checkin.service                   3.285e-01  9.137e-03   35.950  < 2e-16 ***
## Cleanliness                       9.494e-02  1.298e-02    7.312 2.63e-13 ***
## Online.boarding                   1.853e-01  1.258e-02   14.721  < 2e-16 ***
## Departure.Delay.in.Minutes        4.055e-03  1.046e-03    3.876 0.000106 ***
## Arrival.Delay.in.Minutes         -8.583e-03  1.030e-03   -8.337  < 2e-16 ***
## Customer.TypeLoyal Customer       2.466e+00  3.498e-02   70.488  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 131606  on 95403  degrees of freedom
## Residual deviance:  59213  on 95380  degrees of freedom
## AIC: 59261
##
## Number of Fisher Scoring iterations: 6
```

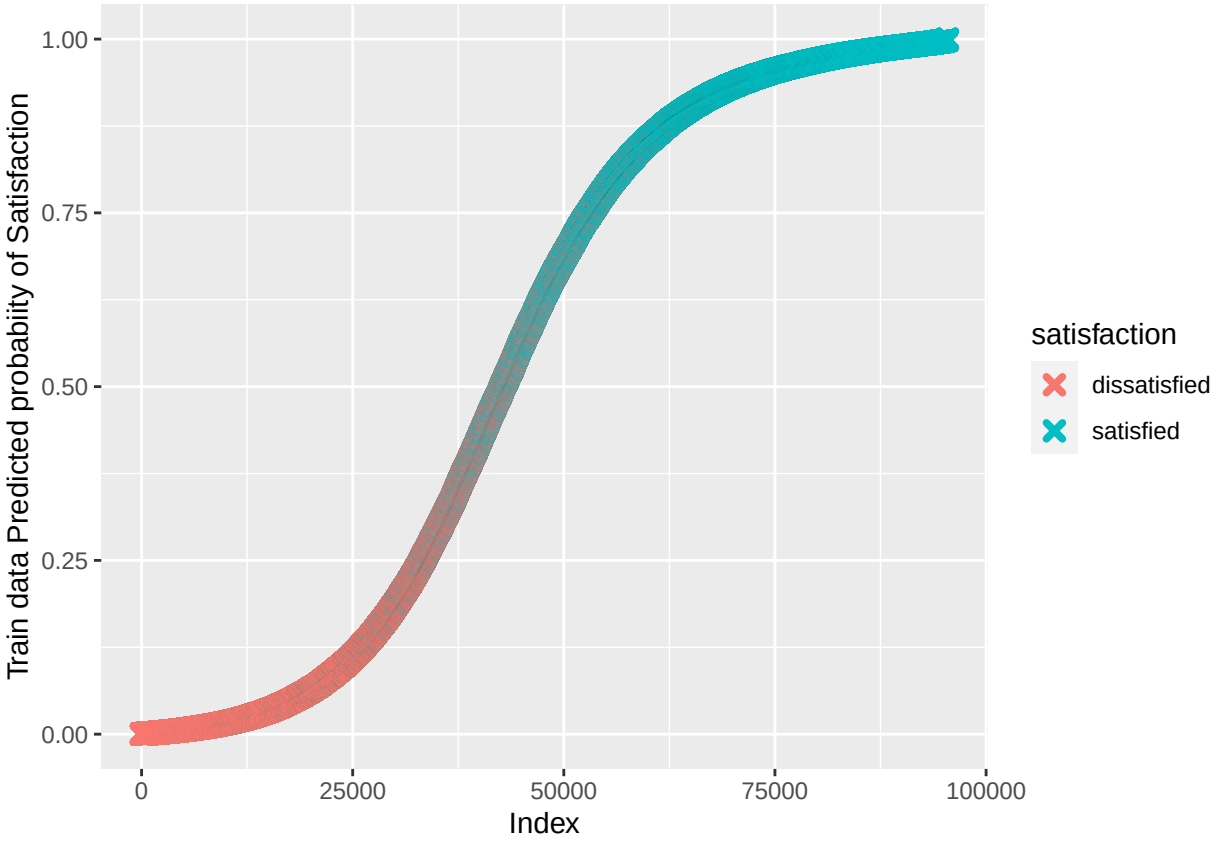
From the results above, we can see that all the P-values are well below the 0.05 significant level. Thus, the log(odds) and log(odds ratios) are statistically signification. Moreover, variable Gender, Age, Type of travel, class, Flight Distance, Gate location, Inflight WiFi service, Departure Arrival time convenience and Arrival Delay negatively contribute to customer satisfaction. Since all the variables are statistically significant, we decided to assess the relative importance of individual predictors in the model.

```
##                                    Overall
## GenderMale                        43.815166
## Age                                4.678556
## Type.of.TravelPersonal Travel     33.459809
## ClassEco                          25.582155
## ClassEco Plus                     21.232120
## Flight.Distance                    8.189018
## Seat.comfort                      46.801706
## Departure.Arrival.time.convenient 37.603884
## Food.and.drink                    14.201082
## Gate.location                      4.998596
## Inflight.wifi.service              8.223310
## Inflight.entertainment            76.109970
## Online.support                     7.848005
## Ease.of.Online.booking            18.972938
## On.board.service                  33.864041
## Leg.room.service                  30.526183
## Baggage.handling                  10.902551
## Checkin.service                   35.950081
## Cleanliness                        7.312242
## Online.boarding                   14.720988
## Departure.Delay.in.Minutes         3.876084
## Arrival.Delay.in.Minutes           8.337013
## Customer.TypeLoyal Customer       70.487757
```

From the individual variable importance test result above, we can see variables “Inflight. entertainment” and “Customer.Type” are highly contributing towards explaining the customer satisfaction rate. The t-statistic for each model parameter from the result indicates that all model variables are significantly different from zero. Hence we choose this model to move on with the further analysis.

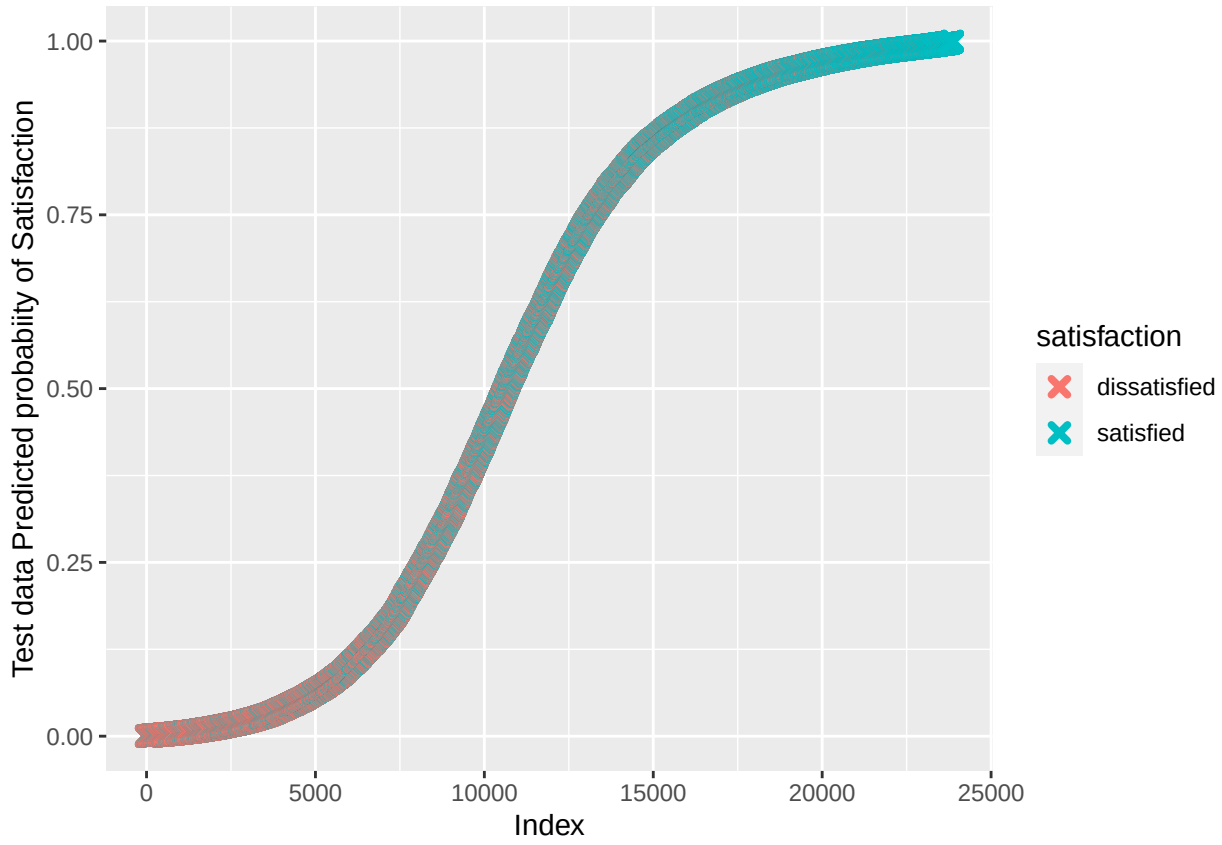


The above graph shows the distribution of fitted values on the training dataset. Then, i plotted the graph showing the predicted probabilities of customer satisfaction and their actual satisfaction.



As we can see from the graph above, most of the satisfied customers are predicted to have a high probability. In contrast, dissatisfied customers are predicted to have a low probability and little overlap in the middle. Then, we proceed to test the model using the test dataset created above.

The results from the first iteration are shown below, with the confusion matrix and accuracy calculation:



The graph above shows the predicted probabilities of customer satisfaction and their actual satisfaction. There is a lot of overlap in the prediction of the test data.

```
##               actual_val
## Predict        dissatisfied satisfied
##   dissatisfied         9173      1471
##   satisfied            1751     11456
```

```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     dissatisfied satisfied
##   dissatisfied         9173      1471
##   satisfied            1751     11456
##
##                Accuracy : 0.8649
##                  95% CI : (0.8605, 0.8692)
##     No Information Rate : 0.542
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7274
##
##  Mcnemar's Test P-Value : 8.869e-07
##
##             Sensitivity : 0.8397
##             Specificity : 0.8862
```

```
##          Pos Pred Value : 0.8618
##          Neg Pred Value : 0.8674
##              Prevalence : 0.4580
##          Detection Rate : 0.3846
##    Detection Prevalence : 0.4463
##       Balanced Accuracy : 0.8630
##
##        'Positive' Class : dissatisfied
##
```

```
## However the missclassification rate on logistic regression model in the first iteration is : 0.135088
```

From the confusion matrix, we can see that the actual dissatisfaction prediction rate is approximately 84%. At the same time, the actual satisfaction rate predicted in 88%. So far, the model is doing well, so we have decided to run the cross-validation on this model.

Stratified K-Fold Cross validation

```
##       Fold Misscalssification_rate
## 1   Fold01                 0.1395
## 2   Fold02                 0.1394
## 3   Fold03                 0.1405
## 4   Fold04                 0.1381
## 5   Fold05                 0.1400
## 6   Fold06                 0.1386
## 7   Fold07                 0.1380
## 8   Fold08                 0.1387
## 9   Fold09                 0.1385
## 10  Fold10                 0.1376
## 11 Average                 0.1389
```

```
## Logistic regression all explanatory variable average of the misclassification rate is : 0.1388882
```

From the Stratified K-Fold cross-validation results, we can see the miss classification rate of prediction in each fold resulted in the range of 13.8 % to 14%, which was close to the miss classification rate in our first iteration. So, we checked the goodness of fit test on the model.

Goodness of fit test

To calculate the r2 values of the model, we have used McFadden pseudo-r2:

```
## fitting null model for pseudo-r2
```

```
##           llh       llhNull            G2      McFadden          r2ML
## -2.960636e+04 -6.580321e+04  7.239370e+04  5.500773e-01  5.317777e-01
##          r2CU
##  7.106603e-01
```

```
## Fitting null model for pseudo-r2 :  0.5500773
```

Using McFadden's R^2 test, which is defined as $1 - \frac{ln(L_M)}{ln(L_0)}$ where $ln(L_M)$ is the log-likelihood value for the fitted model and $ln(L_0)$ is the log-likelihood for the null model with only an intercept as a predictor. The measure values returned were only 55%, indicating that the model lacks predictive power.

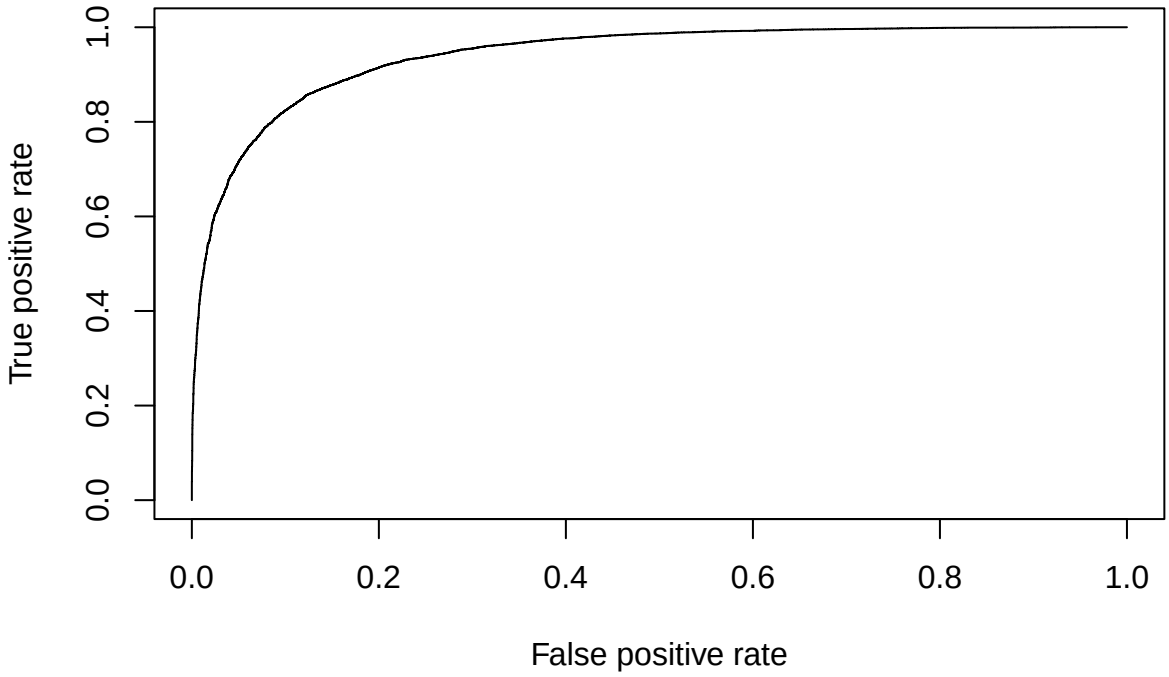
To calculate the P-values of the model, we have used Chi-square test and Hosmer-Lemeshow Test:

```
## P-value from Chi-square test:  0
```

```
## P-value from Hosmer-Lemeshow test:  0
```

Both Chi-square and Hosmer-Lemeshow Test(The Hosmer-Lemeshow test examines whether the observed proportion of events is similar to the predicted probabilities of occurrences in subgroups of the dataset) returned a p-value equal to 0, indicating the model is not a good fit.

Then we calculated the area under ROC curve (receiver operating characteristic curve):




```
## [1] 0.9409821
```

From the result, we can see the proportion of dissatisfied data points that are correctly estimated correct $TPR = \frac{TP}{n(Y=1)}$, and the proportion of satisfied data points that are correctly estimated correctly $TNR = \frac{TN}{n(Y=0)}$ is only 73.96% is below 80% indicating that the model does not do a great job in discriminating between the two categories which comprise our target variable.

Multicollinearity Assumption verification.

Finally, we have checked the condition of multicollinearity:

```
##                                       GVIF Df GVIF^(1/(2*Df))
## Gender                            1.070025  1        1.034420
## Age                               1.216746  1        1.103062
## Type.of.Travel                    1.990953  1        1.411011
## Class                             1.732968  2        1.147355
## Flight.Distance                   1.237354  1        1.112364
## Seat.comfort                      1.967135  1        1.402546
## Departure.Arrival.time.convenient 2.417992  1        1.554989
## Food.and.drink                    2.708128  1        1.645639
## Gate.location                     2.043820  1        1.429622
## Inflight.wifi.service             1.845869  1        1.358628
## Inflight.entertainment            1.307170  1        1.143315
## Online.support                    1.887762  1        1.373959
## Ease.of.Online.booking            2.869400  1        1.693930
## On.board.service                  1.614235  1        1.270525
## Leg.room.service                  1.226103  1        1.107295
## Baggage.handling                  1.820208  1        1.349151
## Checkin.service                   1.187413  1        1.089685
## Cleanliness                       1.949458  1        1.396230
## Online.boarding                   2.196277  1        1.481984
## Departure.Delay.in.Minutes       14.374959  1        3.791432
## Arrival.Delay.in.Minutes         14.392358  1        3.793726
## Customer.Type                     1.399016  1        1.182800
```

From the multicollinearity test results above, all the variables have values below 5, indicating no or little multicollinearity observed between the predictors.

To conclude, all models are wrong, but some are useful; however, based on the results from the goodness of fit, the above model could not be more helpful.

LDA and QDA

Based on the results of generalized linear model, we already know that all the predictors have the p-value smaller than 0.05, indicating all the predictors are significant and should be involved into our model building process. Therefore, we involved all the predictors for both LDA and QDA models

First, we built the model and evaludated the mis-classification rates without the cross validation.

```
##
##                dissatisfied satisfied
##   dissatisfied         9029      1386
##   satisfied            1895     11541
```

```
## [1] 0.1375624
```

```
##
##                dissatisfied satisfied
##   dissatisfied         9196      1960
##   satisfied            1728     10967
```

```
## [1] 0.1546266
```

In order to get more accurate evaluation results, the cross validation was also applied to the LDA and QDA model.

```
## $Fold01
## [1] 0.1365415
##
## $Fold02
## [1] 0.1348564
##
## $Fold03
## [1] 0.136021
##
## $Fold04
## [1] 0.135434
##
## $Fold05
## [1] 0.1362154
##
## $Fold06
## [1] 0.1360663
##
## $Fold07
## [1] 0.1352104
##
## $Fold08
## [1] 0.1360757
##
## $Fold09
## [1] 0.1355086
##
## $Fold10
## [1] 0.1359173
```

```
## [1] "The mean mis-classification rate of lda"
```

```
## [1] 0.1357847
```

```
## $Fold01
## [1] 0.1570111
##
## $Fold02
## [1] 0.1546087
##
## $Fold03
## [1] 0.1570312
##
## $Fold04
## [1] 0.1604972
##
## $Fold05
## [1] 0.1585205
```

```
##
## $Fold06
## [1] 0.1504332
##
## $Fold07
## [1] 0.157348
##
## $Fold08
## [1] 0.1544675
##
## $Fold09
## [1] 0.1563324
##
## $Fold10
## [1] 0.1538247
```

```
## [1] "The mean mis-classification rate of qda"
```

```
## [1] 0.1560074
```

From the results above, the misclassification rate of LDA is around 13%. In detail, the mis classification rates of LDA are 13.75624% for the evaluation results without the cross validation, and 13.57847% for the evaluation results with the cross validation.

The misclassification rate of LDA is around 15% . 15.46266% for the evaluation results without the cross validation, and 15.60074% for the evaluation results with the cross validation.

Therefore, the misclassification rate for LDA is lower so we would choose LDA rather than QDA. However, the assumption of LDA is strict and then we would check the assumption of the equal variance and normality.

$$H_0 \text{ :the data have equal variance}$$
$$H_a \text{ : the data do NOT have equal variance}$$

```
## [1] "Age_Equality of Variance"
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Age by satisfaction
## Bartlett's K-squared = 997.36, df = 1, p-value < 2.2e-16
```

```
## [1] "Flight.Distance_Equality of Variance"
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Flight.Distance by satisfaction
## Bartlett's K-squared = 3220, df = 1, p-value < 2.2e-16
```

```
## [1] "Seat.comfort_Equality of Variance"
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Seat.comfort by satisfaction
## Bartlett's K-squared = 6644.2, df = 1, p-value < 2.2e-16

## [1] "Departure.Arrival.time.convenient_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Departure.Arrival.time.convenient by satisfaction
## Bartlett's K-squared = 60.666, df = 1, p-value = 6.762e-15

## [1] "Food.and.drink_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Food.and.drink by satisfaction
## Bartlett's K-squared = 1423.9, df = 1, p-value < 2.2e-16

## [1] "Gate.location_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Gate.location by satisfaction
## Bartlett's K-squared = 996.37, df = 1, p-value < 2.2e-16

## [1] "Inflight.wifi.service_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Inflight.wifi.service by satisfaction
## Bartlett's K-squared = 588.09, df = 1, p-value < 2.2e-16

## [1] "Inflight.entertainment_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Inflight.entertainment by satisfaction
## Bartlett's K-squared = 343.9, df = 1, p-value < 2.2e-16

## [1] "Online.support_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Online.support by satisfaction
## Bartlett's K-squared = 1472.6, df = 1, p-value < 2.2e-16
```

```
## [1] "Ease.of.Online.booking_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Ease.of.Online.booking by satisfaction
## Bartlett's K-squared = 4456.6, df = 1, p-value < 2.2e-16

## [1] "Leg.room.service_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Leg.room.service by satisfaction
## Bartlett's K-squared = 1477.3, df = 1, p-value < 2.2e-16

## [1] "Baggage.handling_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Baggage.handling by satisfaction
## Bartlett's K-squared = 237.66, df = 1, p-value < 2.2e-16

## [1] "Checkin.service_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Checkin.service by satisfaction
## Bartlett's K-squared = 593.12, df = 1, p-value < 2.2e-16

## [1] "Cleanliness_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Cleanliness by satisfaction
## Bartlett's K-squared = 332.23, df = 1, p-value < 2.2e-16

## [1] "Online.boarding_Equality of Variance"

##
##  Bartlett test of homogeneity of variances
##
## data:  Online.boarding by satisfaction
## Bartlett's K-squared = 1689.6, df = 1, p-value < 2.2e-16

## [1] "Departure.Delay.in.Minutes_Equality of Variance"
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Departure.Delay.in.Minutes by satisfaction
## Bartlett's K-squared = 4767.1, df = 1, p-value < 2.2e-16
```

```
## [1] "Arrival.Delay.in.Minutes_Equality of Variance"
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Arrival.Delay.in.Minutes by satisfaction
## Bartlett's K-squared = 4555.7, df = 1, p-value < 2.2e-16
```

Based on the results above, we got P-value smaller than 0.05, therefore we would reject the null hypothesis. So, the assumption of the equal variance is not met. We also try F test, Levene's test, all tests show the assumption of equal variance is not met.

$$H_0 \text{ :the data follow normal distribution}$$
$$H_a \text{ : the data do NOT follow normal distribution}$$

```
## [1] "Age_Normality"
```

```
## [1] 3.7e-24
```

```
## [1] "Seat.comfort_Normality"
```

```
## [1] 3.7e-24
```

```
## [1] "Departure.Arrival.time.convenient_Normality"
```

```
## [1] 3.7e-24
```

```
## [1] "Food.and.drink_Normality"
```

```
## [1] 3.7e-24
```

```
## [1] "Gate.location_Normality"
```

```
## [1] 3.7e-24
```

```
## [1] "Inflight.wifi.service_Normality"
```

```
## [1] 3.7e-24
```

```
## [1] "Inflight.entertainment_Normality"
```

```
## [1] 3.7e-24
```

```
## [1] "Online.support_Normality"

## [1] 3.7e-24

## [1] "Ease.of.Online.booking_Normality"

## [1] 3.7e-24

## [1] "Leg.room.service_Normality"

## [1] 3.7e-24

## [1] "Baggage.handling_Normality"

## [1] 3.7e-24

## [1] "Checkin.service_Normality"

## [1] 3.7e-24

## [1] "Cleanliness_Normality"

## [1] 3.7e-24

## [1] "Online.boarding_Normality"

## [1] 3.7e-24

## [1] "Departure.Delay.in.Minutes_Normality"

## [1] 3.7e-24

## [1] "Arrival.Delay.in.Minutes_Normality"

## [1] 3.7e-24
```

Based on the results above, we got P-value smaller than 0.05, therefore we would reject the null hypothesis. So, the assumption of the normality is not met.

However, even those assumptions are not met. As we can see from the results of the cross validation with k=10, all the 10 folds gave the misclassification rates at around 13% (13.4% to 13.6%), indicating the misclassfication rates of the LDA model is very stable. So that we could say even the LDA model does not meet the assumption, it still has relevant low misclassfication rate and the stable performance.

Decision Tree

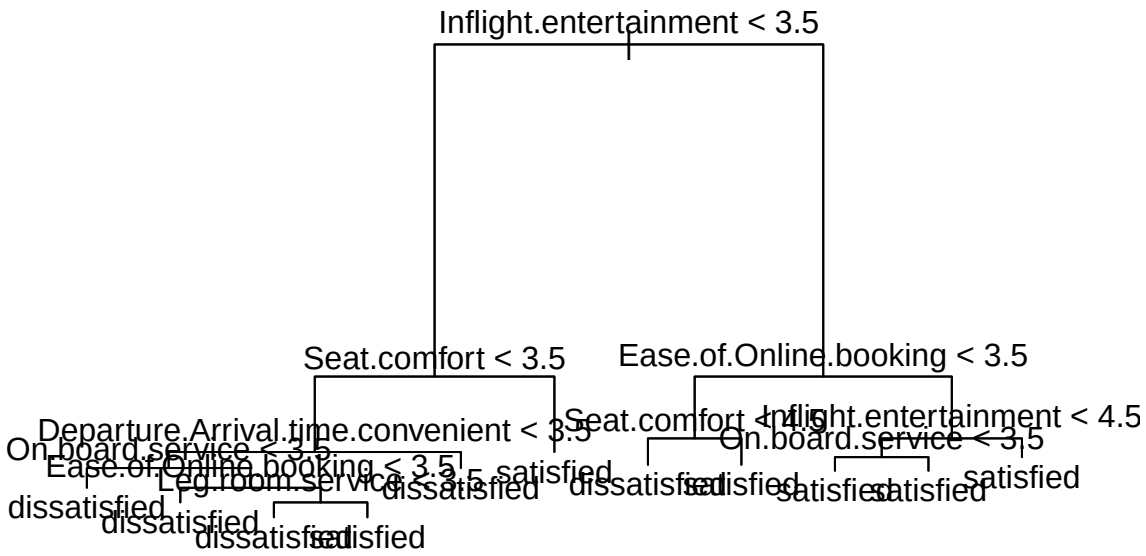
A decision tree is a non-parametric supervised greedy learning algorithm, which is often used for classification or regression problems. In R, we have several ways to construct the tree model, two common packages to construct trees: "tree" package and "rpart" package. "Rpart" offers more flexibility when growing trees. And it is more recommended[3]. It uses the CART algorithm. In this section, I will experiment by using both packages to construct the tree models and try to compare the classification accuracy of the tree constructed by the two packages.

Decision Tree Model Building Using 'tree' Package:

In this part, I will use 'tree' package to construct and train airline customer satisfaction prediction model. We apply the classification tree to the training part and model the relationship between satisfaction and all explanatory variables. The results are shown below:

```
##
## Classification tree:
## tree(formula = factor(satisfaction) ~ ., data = train)
## Variables actually used in tree construction:
## [1] "Inflight.entertainment"            "Seat.comfort"
## [3] "Departure.Arrival.time.convenient" "On.board.service"
## [5] "Ease.of.Online.booking"            "Leg.room.service"
## Number of terminal nodes:  11
## Residual mean deviance:  0.6332 = 60400 / 95390
## Misclassification error rate: 0.1393 = 13286 / 95404
```

Then we plot the tree. From the dendrogram and summary information we can see "Inflight.entertainment", "Seat.comfort", "Ease.of.Online.booking", "On.board.service", "Leg.room.service", "Departure.Arrival.time.convenient" are used to construct the unpruned tree based on the top-down greedy method.

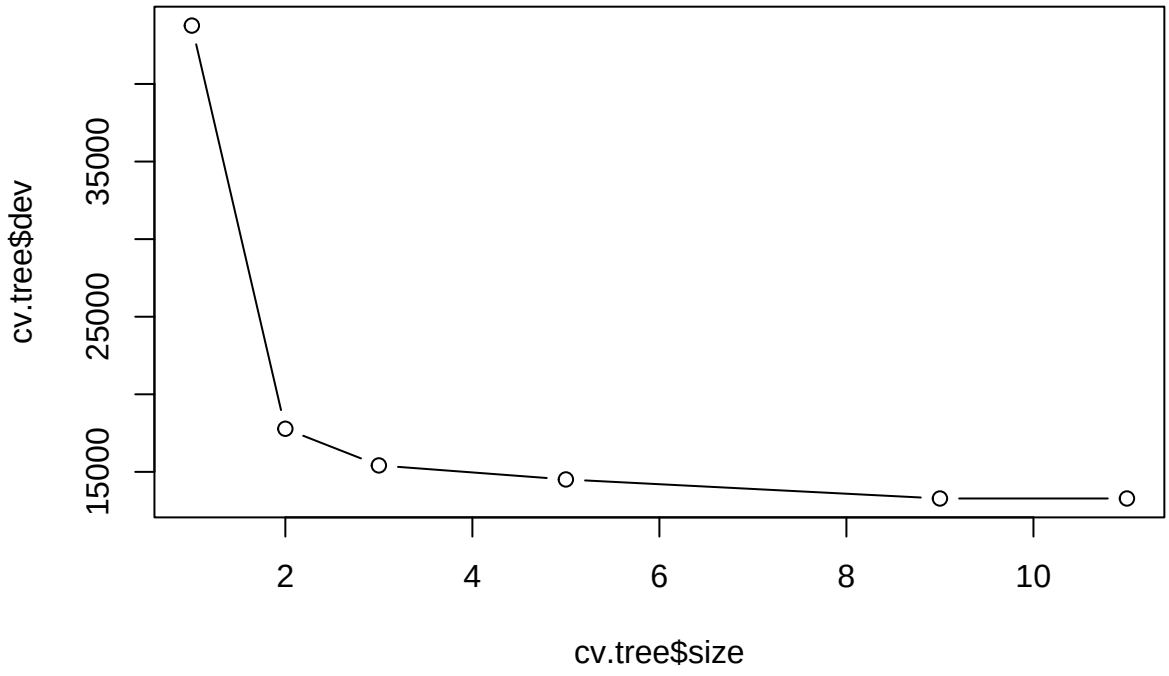


We can use the unpruned tree model in the test set to predict whether the passengers are satisfied with their airline. The misclassification rate is 13.84%.

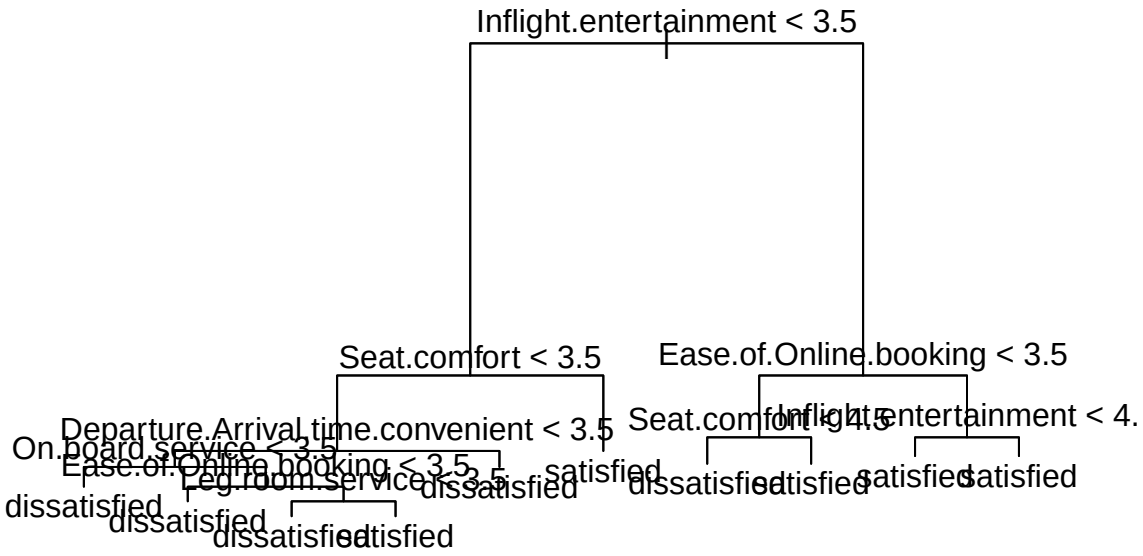
```
##
```

```
## satisfaction.test dissatisfied satisfied
##      dissatisfied         9354      1731
##      satisfied            1570     11196
```

In order to prevent model overfitting and control the complexity of the tree, we want to use the cross validation to prune the tree and select the best number of terminal nodes. From the below graph, we can find the minimum cross-validation error occurs at a tree size of 9, so, I choose tree size of 9.



We then prune the tree and visualize the pruned tree:



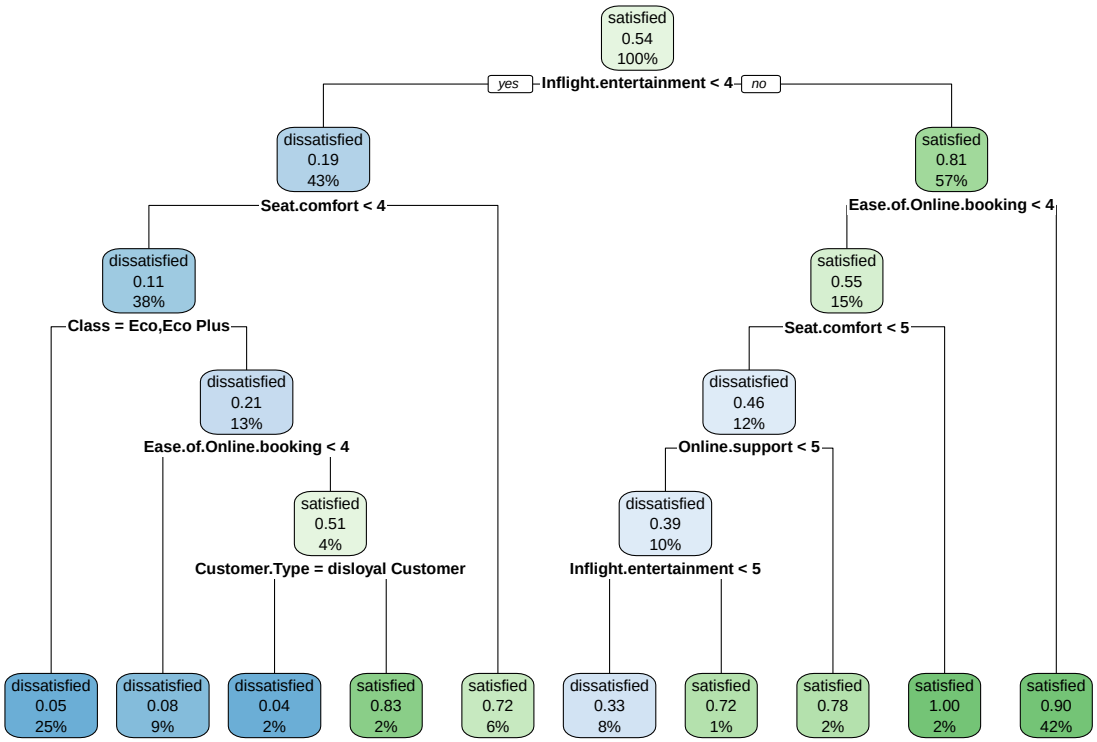
We then use the new pruned tree model in the test set to predict whether the passengers are satisfied with their airline. The misclassification rate now is still 13.84%. Compared to unpruned tree, it reduced two terminal nodes.

```
##
## tree.prune     dissatisfied satisfied
##   dissatisfied         9354      1731
##   satisfied            1570     11196
```

Decision Tree Model Building Using 'rpart' Package:

In this part, I will use 'rpart' package to construct and train airline customer satisfaction prediction model. First, visualize the tree model. From the dendrogram we can see "Inflight.entertainment", "Seat.comfort", "Ease.of.Online.booking", "Online.support", "Class", "Customer Type" are used to construct the tree.

Compared with 'tree' package, the model built using 'rpart' package has different variable selection strategy. Two trees are different, they use different variables to construct the tree.



We can use this model in the test set to predict whether the passengers are satisfied with their airline. The misclassification rate here is 11.86%. Compared with 'tree' package, the model using 'rpart' gives more accurate result.

```
##
## tree_rpart     dissatisfied satisfied
##   dissatisfied         9243      1147
##   satisfied            1681     11780
```

K-fold Stratified Cross Validation for Classification Tree

Now we use cross validation to test the model performance. Here I test the tree model constructed by using the "tree" package, and select the stratified cross validation with K value of 10. Through testing, we find that the tree model is obtained the average misclassification rate of 13.91% (both unpruned and pruned tree).

```
## $Fold01
## [1] 0.1418868
##
## $Fold02
## [1] 0.1410364
##
## $Fold03
## [1] 0.1402817
##
## $Fold04
```

```
## [1] 0.1346638
##
## $Fold05
## [1] 0.1372746
##
## $Fold06
## [1] 0.1365199
##
## $Fold07
## [1] 0.13676
##
## $Fold08
## [1] 0.1451572
##
## $Fold09
## [1] 0.1382693
##
## $Fold10
## [1] 0.1390356
```

```
## [1] "The mean mis-classification rate of unpruned tree (tree package)"
```

```
## [1] 0.1390885
```

```
## $Fold01
## [1] 0.1418868
##
## $Fold02
## [1] 0.1410364
##
## $Fold03
## [1] 0.1402817
##
## $Fold04
## [1] 0.1346638
##
## $Fold05
## [1] 0.1372746
##
## $Fold06
## [1] 0.1365199
##
## $Fold07
## [1] 0.13676
##
## $Fold08
## [1] 0.1451572
##
## $Fold09
## [1] 0.1382693
##
## $Fold10
## [1] 0.1390356
```

```
## [1] "The mean mis-classification rate of pruned tree (tree package)"
```

```
## [1] 0.1390885
```

The average misclassification rate of the tree model built using the "rpart" package is about 11.89%. Compared with 'tree' package, the model using 'rpart' gives more accurate prediction result.

```
## $Fold01
## [1] 0.1190776
##
## $Fold02
## [1] 0.1173067
##
## $Fold03
## [1] 0.1222539
##
## $Fold04
## [1] 0.1179775
##
## $Fold05
## [1] 0.1195807
##
## $Fold06
## [1] 0.1144654
##
## $Fold07
## [1] 0.1149589
##
## $Fold08
## [1] 0.124109
##
## $Fold09
## [1] 0.1203253
##
## $Fold10
## [1] 0.1185744
```

```
## [1] "The mean mis-classification rate of tree (rpart package)"
```

```
## [1] 0.118863
```

In conclusion, tree model construction using 'rpart' package seems gives a more accurate prediction result.

For the airline data set, compare the accuracy of using the cumulative logit model and continuation-ratio logit model.

Conclusion

Overall, we aim to predict customer satisfaction with their travel experience based on the features provided by the airline and the equipment on board the aircraft. This analysis involves exploring a dataset of airline satisfaction, conducting preliminary analyses such as exploratory data analysis, categorical data analysis, classification analysis, and predictive analysis. - The logistic regression Model is not a good fit to predict

customer satisfaction. No multicollinearity is detected between the variables. - The LDA model has the miss-classification rate at 13.13%. The QDA model has the miss-classification rate at 15.48%. - The tree-based analysis using rpart package has miss-classification rate about 11.69%

In conclusion, we would like to choose tree-based analysis using rpart package because it has the lowest Miss-classification rate.

References

[1] Airlines Customer Satisfaction.
https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction. Accessed 27 Jan. 2023. [2] ggplot2 : Quick correlation matrix heatmap - R software and data visualization.
http://sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization#prepare-the-data. Accessed 10 Feb. 2023. [3] Rohitschauhan and N. *, "Home," rohitschauhan, 13-Jun-2018. [Online]. Available:
http://www.rohitschauhan.com/index.php/2018/06/13/a-comparison-on-using-r-tree-vs-r-rpart/#:~:text=Rpart%20offers%20more%20flexibility%20when,mincut%2C%20minsize%20and%20mindev).
[Accessed: 05-Feb-2023].