CS 513 - KNOWLEDGE DISCOVERY & DATA MINING

Airline Passenger Satisfaction

HARSH AGRAWAL
CWID- 10475285
CS-513B

AMIT RAMJEE
CWID-10457040
CS-513B

ABHISHEK AMBERKAR
CWID- 10469370
CS-513B

SHWETA MISHRA
CWID-10476307
CS-513A

Team Members

TODAY'S AGENDA

1. Discussion of Dataset
2. Problem Statement / Objective
3. Overview of Classification Techniques
4. EDA & Correlation Heatmap
5. Detailed Classification/Clustering Analysis
6. Conclusions

AIRPLANE PASSENGER SATISFACTION DATASET

Link: Airline Passenger Satisfaction | Kaggle

- Training data: 25 columns and 103,905 training rows of data.
- Test data: 25 columns and 25,977 rows
- Categorical & Numeric columns
- Binary outcome variable: Satisfaction as 'satisfied' or 'neutral of dissatisfied'
- Passenger numerical ratings (1 – 5)
- Cleaned data : replaced any blanks with median for Arrival Delay in Minutes

Dataset Variables Attributes

Row
id
Gender
Customer Type
Age
Type of Travel
Class
Flight Distance
Inflight wifi service
Departure/Arrival time convenient
Ease of Online booking
Gate location
Food and drink Online boarding
Seat comfort
Inflight entertainment
On-board service
Leg room service
Baggage handling
Checkin service
Inflight service
Cleanliness
Departure Delay in Minutes
Arrival Delay in Minutes
Satisfaction
Business
Eco Plus

PROBLEM STATEMENT & OBJECTIVE

Problem Statement

- Airline passengers often times experience a less than satisfying travel experience despite efforts avoid a bad experience (seat selection, travel class type, Wi-Fi, food options).

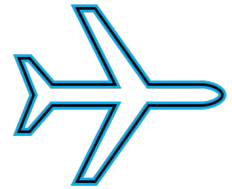
- There is an opportunity for airlines operating in the industry to improve their overall service across the board to help optimize passenger satisfaction, therefore it is key for airlines to understand which factors play a role in this.

Objective

- Predict the satisfaction of an airline passengers based on their respective different taggings and customer feedback.

- To understand the categories (feedback and taggings) which have a larger influence/impact upon passenger satisfaction across the board.

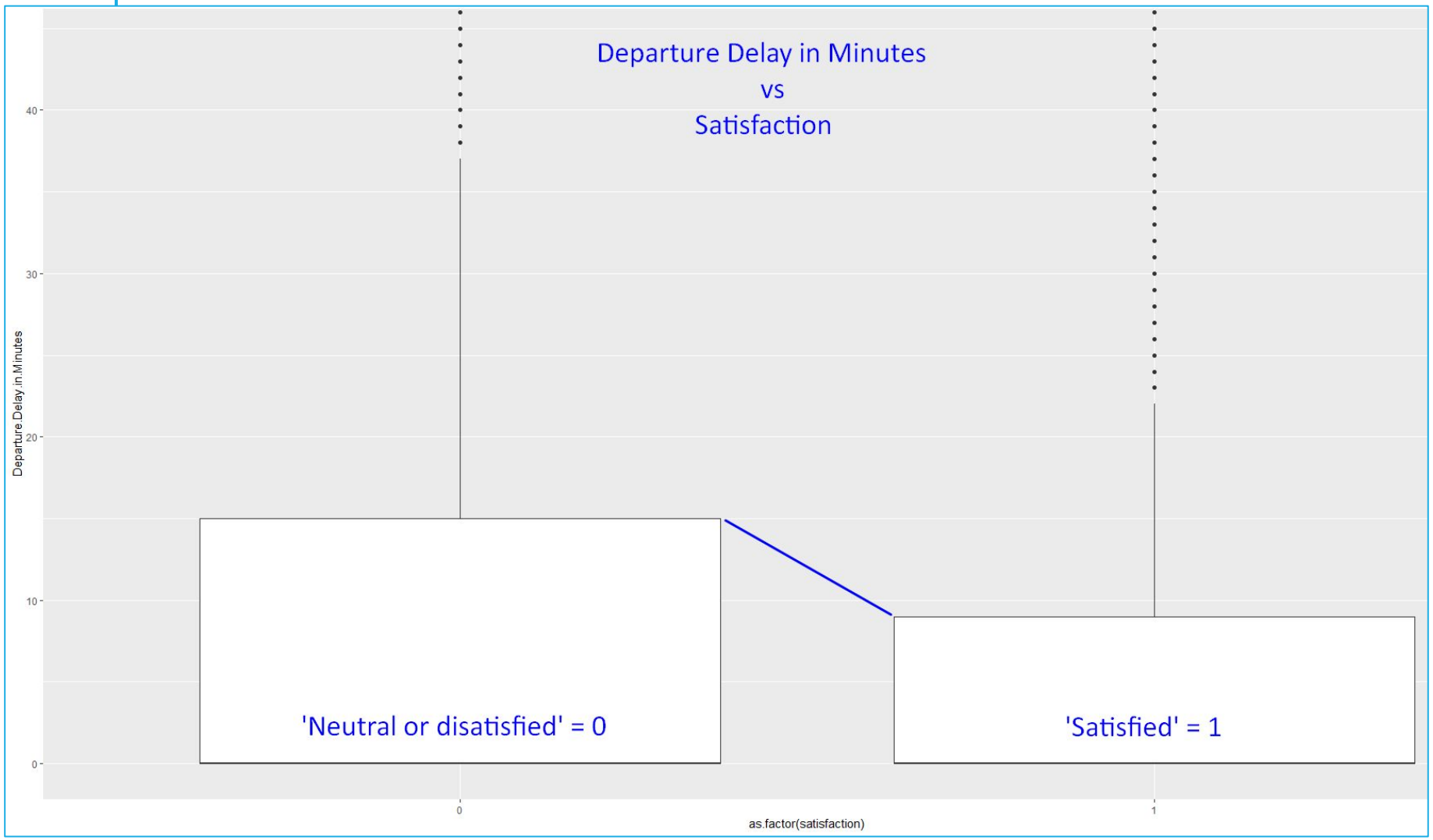
CLASSIFICATION TECHNIQUES USED



Techniques

- Naive Bayes
- KNN
- CART Decision Tree
- C5.0 Algorithm
- Random Forest
- SVM

EDA: DOES DEPARTURE DELAY IMPACT SATISFACTION?



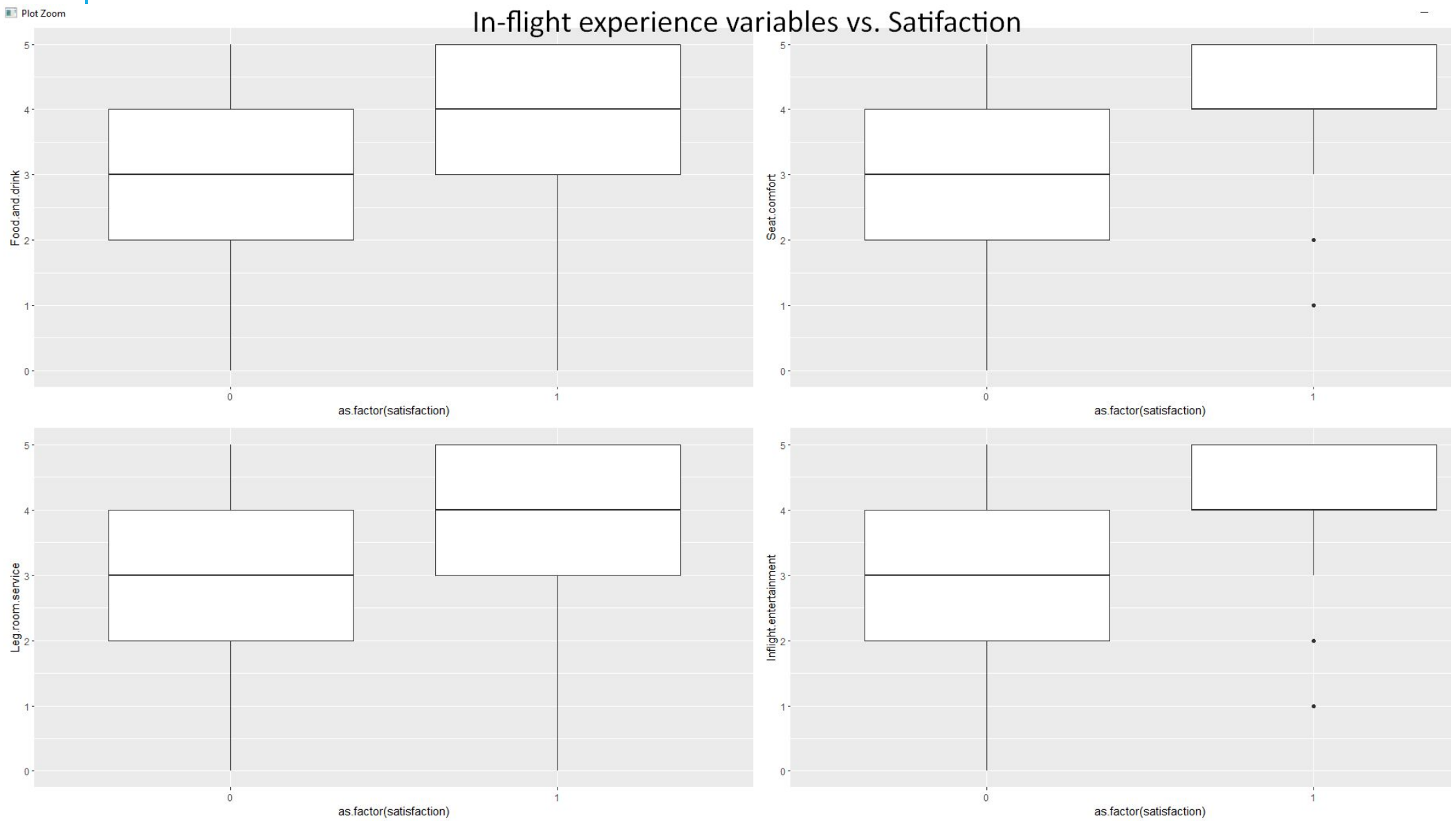
- Shouldn't departure delay have an impact on passenger satisfaction?

- Boxplot shows airplane departure delayal does not impact passenger satisfaction. There is little difference in delayal minutes between those satisfied vs. not (~5-7 minutes)

EDA: HOW DOES AIRLINE CLASS INFLUENCE SATISFACTION?

- As expected, satisfied passengers were more likely to fall in the business class (approx.75%) vs. 'neutral or dissatisfied' who mostly flew Economy class (~60%).

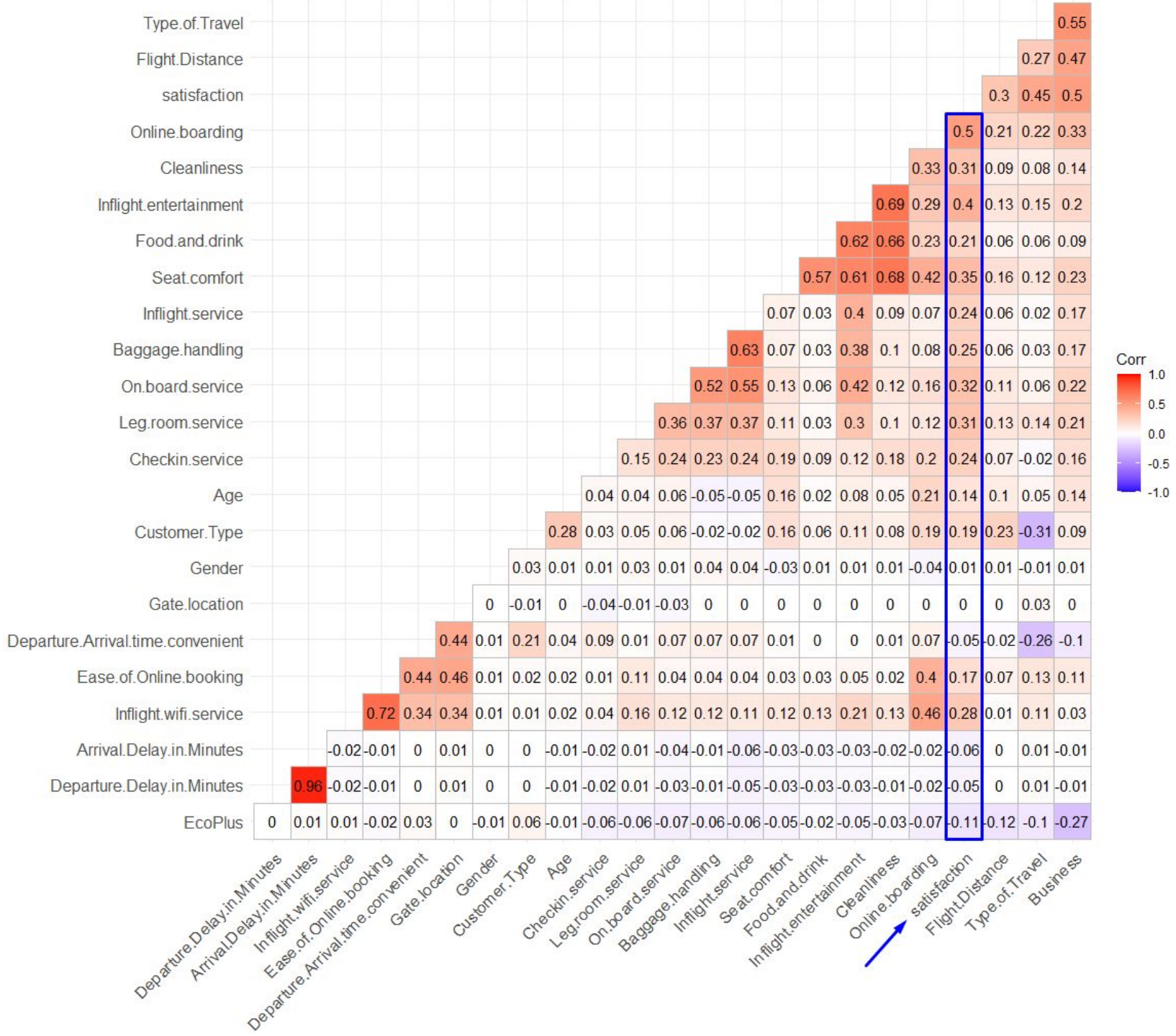
EDA: HOW DOES AIRLINE CLASS INFLUENCE SATISFACTION?



- Satisfied passengers gave one rating better than those who were ‘neutral or dissatisfied’.

CORRELATION HEATMAP

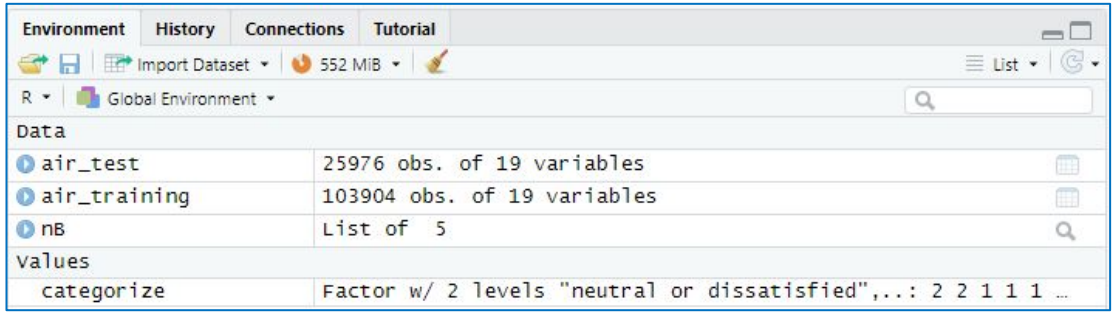
- Changed categorical variables into numeric, split the Airline class into Business vs. Eco Plus
- Selected attributes with higher correlation to satisfaction, which was mostly in-flight passenger experience



NAÏVE BAYES

- Accuracy: ~85.4%

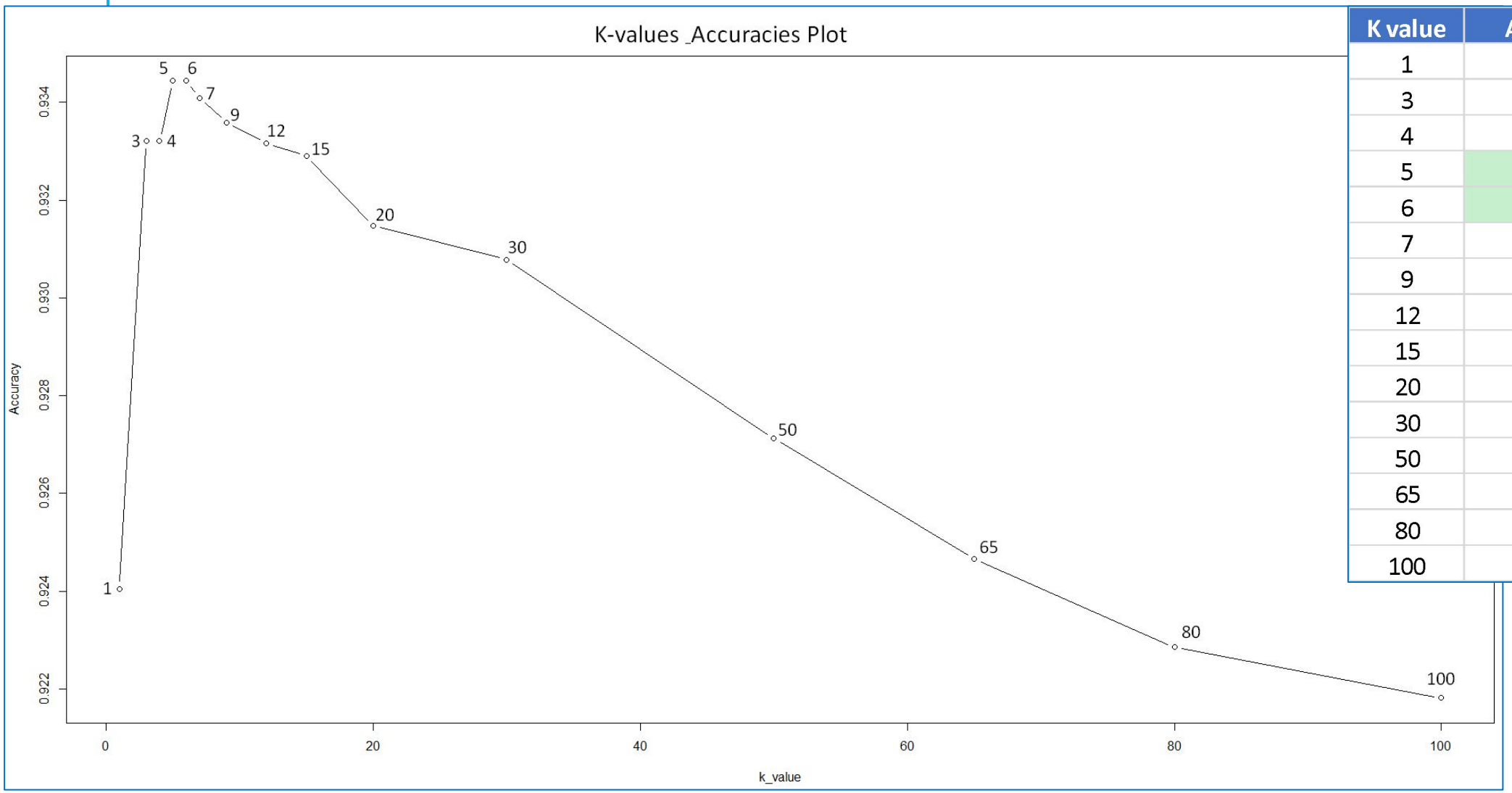
```
Confusion Matrix and Statistics

                          categorize
                           neutral or dissatisfied satisfied
 neutral or dissatisfied                    12942      1631
 satisfied                                   2151      9252

               Accuracy : 0.8544
                 95% CI : (0.8501, 0.8587)
    No Information Rate : 0.581
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7029

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8575
            Specificity : 0.8501
         Pos Pred Value : 0.8881
         Neg Pred Value : 0.8114
             Prevalence : 0.5810
         Detection Rate : 0.4982
   Detection Prevalence : 0.5610
      Balanced Accuracy : 0.8538

       'Positive' Class : neutral or dissatisfied
```



KNN METHOD

- Accuracy: ~93.4%

```
# Code to run the KNN model for multiple K values:
k_value<-c(1,3,4,5,6,7,9,12,15,20,30,50,65,80,100) # K value to test
error_Rate<-vector(length = length(k_value))  # store error rate for each model

library(kknn) # use library to run kknn formula
for(i in 1:length(k_value)){
  knn_model<-kknn(satisfaction~., air_training, air_test,k = k_value[i],kernel = "rectangular")
  predicted<-fitted(knn_model)
  wrong<-sum(predicted!=air_test$satisfaction) # count number of wrong where test data doesn't prediction
  error_Rate[i]<-wrong/length(predicted)
}
print(1-error_Rate) # Print Accuracy
plot(k_value, 1-error_Rate, type = "b", ylab = "Accuracy") # Plot the accuracy
```

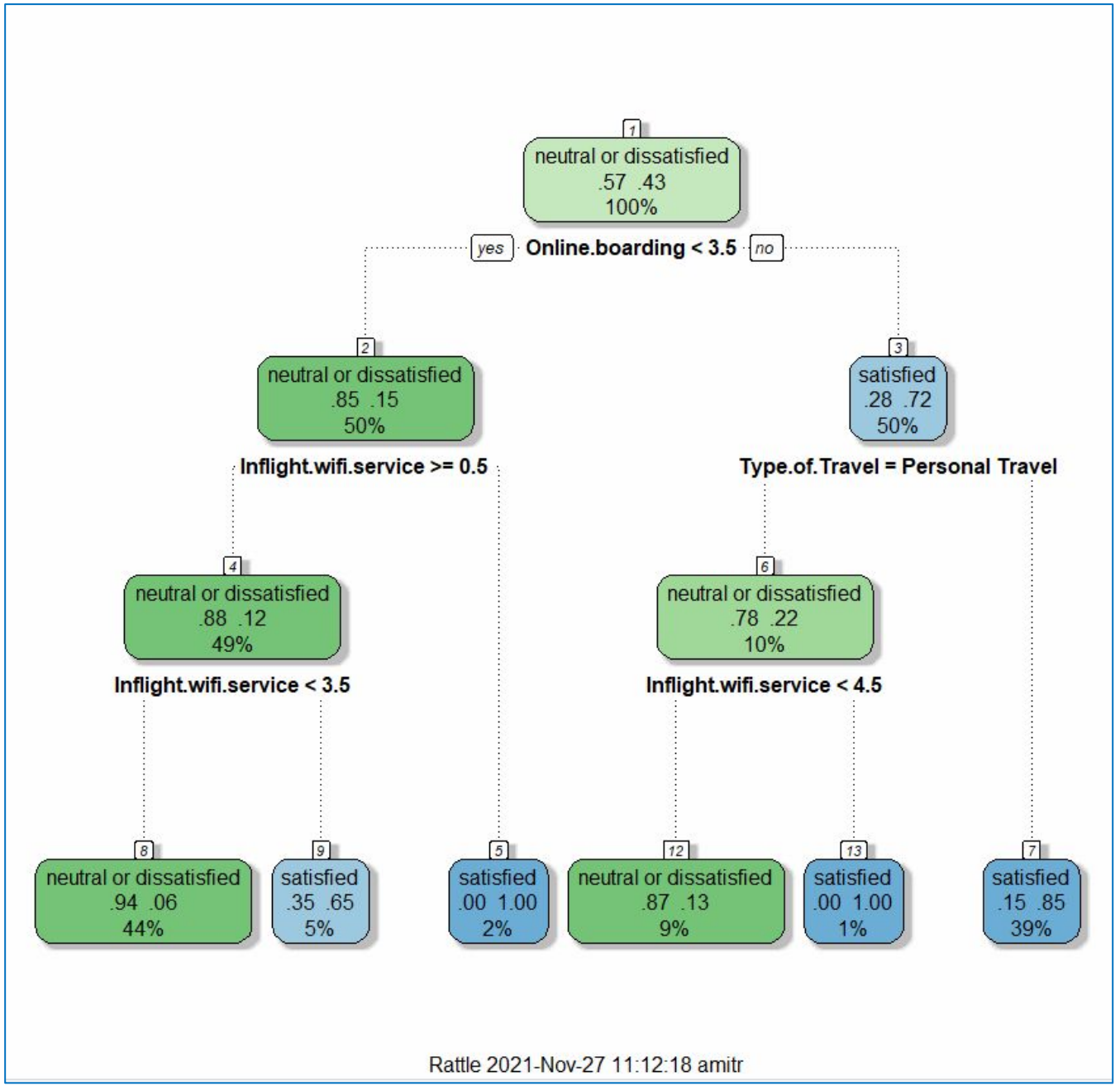


K value	Accuracy	Error Rate
1	0.92404530	0.07595470
3	0.93320760	0.06679240
4	0.93320760	0.06679240
5	0.93443950	0.06556050
6	0.93443950	0.06556050
7	0.93409300	0.06590700
9	0.93359250	0.06640750
12	0.93316910	0.06683090
15	0.93289960	0.06710040
20	0.93147520	0.06852480
30	0.93078230	0.06921770
50	0.92712500	0.07287500
65	0.92466120	0.07533880
80	0.92285190	0.07714810
100	0.92181240	0.07818760

CART DECISION TREE

- Accuracy: ~88.3%

```
Confusion Matrix and Statistics

                          prediction
actual                     neutral or dissatisfied satisfied
  neutral or dissatisfied                   12599      1974
  satisfied                                  1045     10358

               Accuracy : 0.8838
                 95% CI : (0.8798, 0.8877)
    No Information Rate : 0.5253
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7661

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9234
            Specificity : 0.8399
         Pos Pred Value : 0.8645
         Neg Pred Value : 0.9084
             Prevalence : 0.5253
         Detection Rate : 0.4850
   Detection Prevalence : 0.5610
      Balanced Accuracy : 0.8817

       'Positive' Class : neutral or dissatisfied
```



C5.0 ALGORITHM

- Accuracy: ~96.0%,

```
Confusion Matrix and Statistics

                          C50
actual                     neutral or dissatisfied satisfied
  neutral or dissatisfied                    14250       323
  satisfied                                    711     10692

               Accuracy : 0.9602
                 95% CI : (0.9577, 0.9625)
    No Information Rate : 0.576
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9189

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9525
            Specificity : 0.9707
         Pos Pred Value : 0.9778
         Neg Pred Value : 0.9376
             Prevalence : 0.5760
         Detection Rate : 0.5486
   Detection Prevalence : 0.5610
      Balanced Accuracy : 0.9616

       'Positive' Class : neutral or dissatisfied
```

Training Data Summary

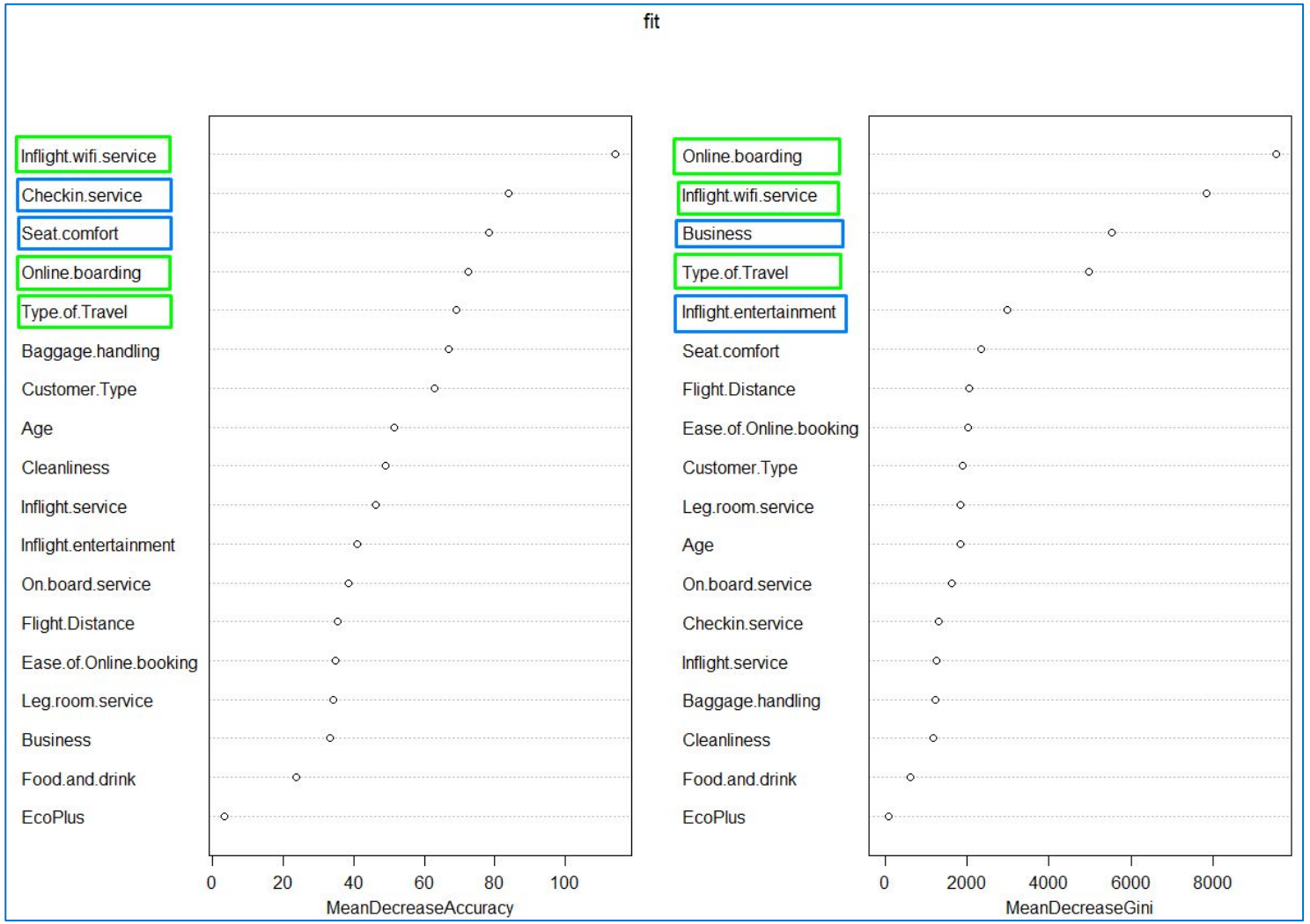
```
Attribute usage:

100.00% Inflight.wifi.service
 91.89% Online.boarding
 78.24% Type.of.Travel
 74.27% Business
 49.19% Customer.Type
 46.30% Cleanliness
 43.79% Checkin.service
 41.03% Inflight.service
 38.72% Baggage.handling
 34.67% On.board.service
 34.27% Seat.comfort
 27.20% Food.and.drink
 19.44% Inflight.entertainment
 18.06% Ease.of.Online.booking
  9.93% Age
  4.26% Leg.room.service
  1.54% Flight.Distance
  0.13% EcoPlus
```

RANDOM FOREST

- Accuracy: ~96.2%

```
Confusion Matrix and Statistics

                      Prediction
actual                 neutral or dissatisfied satisfied
  neutral or dissatisfied                14273       300
  satisfied                                672     10731

               Accuracy : 0.9626
                 95% CI : (0.9602, 0.9649)
    No Information Rate : 0.5753
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9238

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9550
            Specificity : 0.9728
         Pos Pred Value : 0.9794
         Neg Pred Value : 0.9411
             Prevalence : 0.5753
         Detection Rate : 0.5495
   Detection Prevalence : 0.5610
      Balanced Accuracy : 0.9639

       'Positive' Class : neutral or dissatisfied
```

Fit Plot



SVM

- Accuracy: ~95.5%

```
Call:
svm(formula = satisfaction ~ ., data = air_training)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  15266
```

Model Summary

```
Confusion Matrix and Statistics

                          svm.pred
actual                     neutral or dissatisfied satisfied
  neutral or dissatisfied                    14127       446
  satisfied                                    723     10680

               Accuracy : 0.955
                 95% CI : (0.9524, 0.9575)

    No Information Rate : 0.5717
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9084

 Mcnemar's Test P-Value : 6.894e-16

            Sensitivity : 0.9513
            Specificity : 0.9599
         Pos Pred Value : 0.9694
         Neg Pred Value : 0.9366
             Prevalence : 0.5717
         Detection Rate : 0.5438
   Detection Prevalence : 0.5610
      Balanced Accuracy : 0.9556

       'Positive' Class : neutral or dissatisfied
```

CONCLUSIONS

We have derived the following conclusions

- Customer satisfaction is most closely linked to their on-flight experience.
- If the outliers are ignored, there is little difference in satisfaction for a flight delay departure of 5 to 7 minutes.
- In comparison to neutral or unsatisfied consumers, satisfied customers provided a higher rating for in-flight user experience factors (seat comfort, food/drink, leg room, entertainment)
- The Random Forest method yielded the highest accuracy for the user's contentment based on their experiences followed closely by the C5.0 and SVM.

Method	Accuracy
NAÏVE BAYES	85.4%
KNN	93.4%
CART DECISION TREE	88.3%
C5.0	96.0%
RANDOM FOREST	96.2%
SVM	95.5%

OTHER INSIGHTS

- Half of the customers travelling in business class were observed to be loyal customers whereas this ratio is almost 10% lower in the economic class. This suggests, more benefits need to be provided to economic class passengers to try convert them into loyal customers.

- Within the Economy class, only 18% of Loyal customers are satisfied, where 34% of regular passengers are satisfied. Therefore, can airlines provide better service/benefits to their regular passenger flying Economy class. On the contrary, Business class passengers that were loyal customer (78%) were more satisfied than regular customers (65%).

Row Labels	Business	Eco	Eco Plus	Grand Total
disloyal Customer	**7,356**	**10,910**	**715**	**18,981**
neutral or dissatisfied	4,447	9,383	659	14,489
satisfied	2,909	1,527	56	4,492
Loyal Customer	**42,309**	**35,835**	**6,779**	**84,923**
neutral or dissatisfied	10,738	28,661	4,991	44,390
satisfied	31,571	7,174	1,788	40,533
Grand Total	**49,665**	**46,745**	**7,494**	**103,904**

Business	Eco	Eco Plus
39%	**57%**	**4%**
31%	65%	5%
65%	34%	1%
50%	**42%**	**8%**
24%	65%	11%
78%	18%	4%
48%	**45%**	**7%**

- Online support appears to be a key determinant in influencing satisfaction. Customer service is a company's opportunity to connect with customers, solve problems, and show they care.

FURTHER ANALYSIS

Thing to take into consideration:

- Names of different airlines would be beneficial to have in analyzing airline passenger satisfaction

- Location of airport boarding could also shed light on customer satisfaction. City location could have a bigger impact than the current columns.

- Conduct a separate study on seat comfort. Seat comfort goes beyond the materials. Separate study to conduct how seat comfort impacts passenger satisfaction.

- 1.1% of the training dataset were customers below the age of 14 were observed to be travelling under “Business Travel”. This was an anomaly in the dataset.