Classification

Isabelle Villegas

2022-09-26

Classification Assignment

This data given by an airline organization. The actual name of the company is not given due to various purposes which is why the name Invistico airlines.

This dataset consists of the details of customers who have already flown with them. The feedback of the customers on various context and their flight data has been consolidated.

The main purpose of this dataset is to predict whether a future customer would be satisfied with their service given the details of the other parameters values.

Also the airlines need to know on which aspect of the services offered by them have to be emphasized more to generate more satisfied customers.

The link for the data set can be found here: https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction (https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction)

```
data <- read.csv("Invistico_Airline.csv")
summary(data)
```

```
##  satisfaction          Gender          Customer.Type           Age
##  Length:129880      Length:129880      Length:129880      Min.   : 7.00
##  Class :character   Class :character   Class :character   1st Qu.:27.00
##  Mode  :character   Mode  :character   Mode  :character   Median :40.00
##                                                           Mean   :39.43
##                                                           3rd Qu.:51.00
##                                                           Max.   :85.00
##
##  Type.of.Travel        Class           Flight.Distance  Seat.comfort
##  Length:129880      Length:129880      Min.   :  50    Min.   :0.000
##  Class :character   Class :character   1st Qu.:1359    1st Qu.:2.000
##  Mode  :character   Mode  :character   Median :1925    Median :3.000
##                                        Mean   :1981    Mean   :2.839
##                                        3rd Qu.:2544    3rd Qu.:4.000
##                                        Max.   :6951    Max.   :5.000
##
##  Departure.Arrival.time.convenient Food.and.drink  Gate.location
##  Min.   :0.000                     Min.   :0.000   Min.   :0.00
##  1st Qu.:2.000                     1st Qu.:2.000   1st Qu.:2.00
```

```
##  Median :3.000                     Median :3.000   Median :3.00
##  Mean   :2.991                     Mean   :2.852   Mean   :2.99
##  3rd Qu.:4.000                     3rd Qu.:4.000   3rd Qu.:4.00
##  Max.   :5.000                     Max.   :5.000   Max.   :5.00
##
##  Inflight.wifi.service Inflight.entertainment Online.support
##  Min.   :0.000         Min.   :0.000          Min.   :0.00
##  1st Qu.:2.000         1st Qu.:2.000          1st Qu.:3.00
##  Median :3.000         Median :4.000          Median :4.00
##  Mean   :3.249         Mean   :3.383          Mean   :3.52
##  3rd Qu.:4.000         3rd Qu.:4.000          3rd Qu.:5.00
##  Max.   :5.000         Max.   :5.000          Max.   :5.00
##
##  Ease.of.Online.booking On.board.service Leg.room.service Baggage.handling
##  Min.   :0.000          Min.   :0.000    Min.   :0.000    Min.   :1.000
##  1st Qu.:2.000          1st Qu.:3.000    1st Qu.:2.000    1st Qu.:3.000
##  Median :4.000          Median :4.000    Median :4.000    Median :4.000
##  Mean   :3.472          Mean   :3.465    Mean   :3.486    Mean   :3.696
##  3rd Qu.:5.000          3rd Qu.:4.000    3rd Qu.:5.000    3rd Qu.:5.000
##  Max.   :5.000          Max.   :5.000    Max.   :5.000    Max.   :5.000
##
##  Checkin.service  Cleanliness    Online.boarding Departure.Delay.in.Minutes
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :   0.00
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:   0.00
##  Median :3.000   Median :4.000   Median :4.000   Median :   0.00
##  Mean   :3.341   Mean   :3.706   Mean   :3.353   Mean   :  14.71
##  3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:  12.00
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :1592.00
##
##  Arrival.Delay.in.Minutes
##  Min.   :   0.00
##  1st Qu.:   0.00
##  Median :   0.00
##  Mean   :  15.09
##  3rd Qu.:  13.00
##  Max.   :1584.00
##  NA's   :393
```

a. Divide into 80/20 train/test

```
split <- round(nrow(data)*0.8)
training <- data[1:split, ]
test <- data[(split+1):nrow(data),]
```

b. Use at least 5 R functions for data exploration, using the training data

Using the mean, max, min, median, and sum functions I am able to find the average, max, min, and median age of the people filling out the surveys for the airline. I also find out the sum of the distances of the flights that were taken by the people filling out the surveys.

```
mean(training$Age)
```

```
## [1] 38.35996
```

```
max(training$Age)
```

```
## [1] 85
```

```
min(training$Age)
```

```
## [1] 7
```

```
median(training$Age)
```

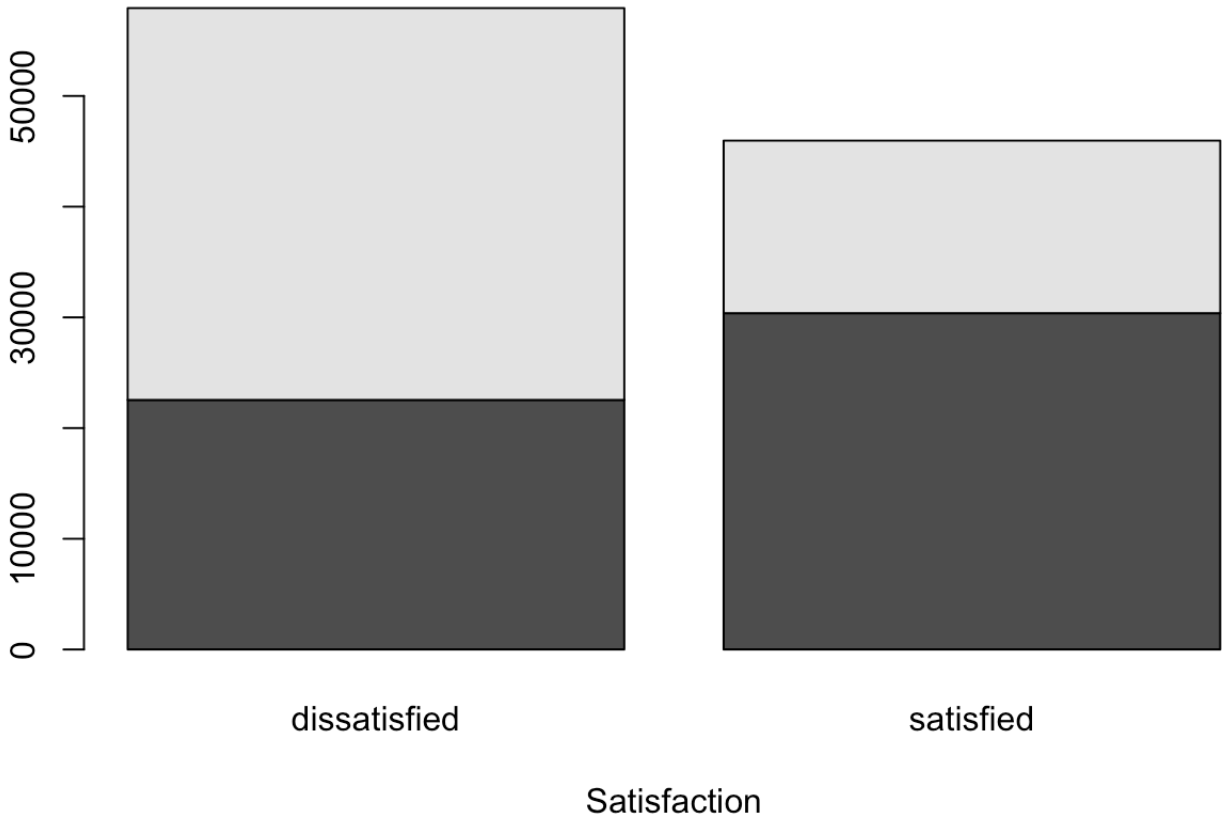
```
## [1] 38
```

```
sum(training$Flight.Distance)
```

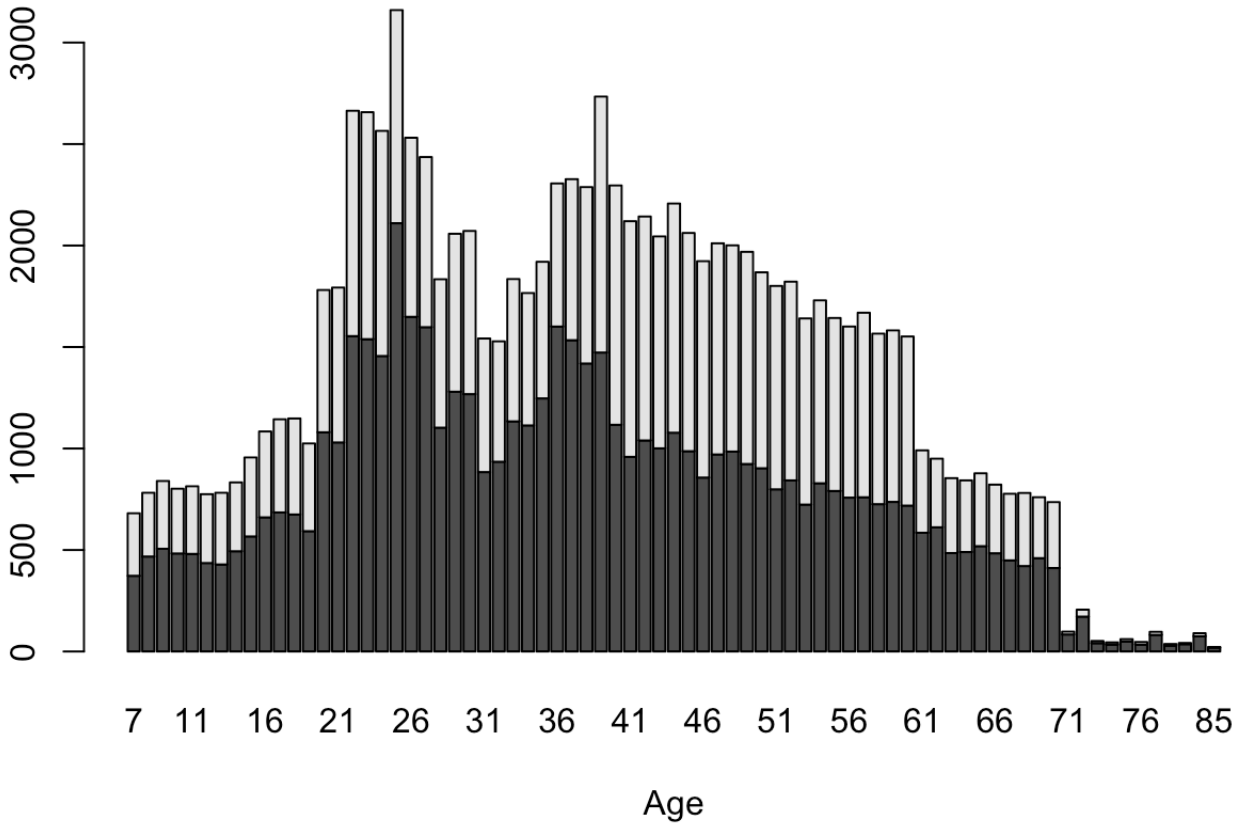
```
## [1] 203579125
```

c. Create at least 2 informative graphs, using the training data

This is a very curious bar plot I was able to create in where Satisfaction was measured between men and women. Women in this data set, the darker color in the plot, were more likely to vocalize their satisfaction than their dissatisfaction. Although, not by very much, and in turn, men, the lighter color, were highly likely to vocalize their dissatisfaction than their satisfaction by much.



In this bar plot, we are able to see that among all the ages that had data on the flight they went on, customers that gave feedback were most likely in their younger ages, around 26. They were also much more likely to vocalize their dissatisfaction than their satisfaction.



d. Build a logistic regression model and output the summary. Write a thorough explanationof the information in the model summary.

Before I am able to run the logit function, I need to factor the satisfaction variable from integer to factor

Because Null Deviance and Residual Deviance have a pretty large gap, it is predicted that the model is a good fit since the difference is big enough.

```
training$satisfaction <- as.factor(training$satisfaction)
glm1 <- glm(satisfaction~Age+Gender+Customer.Type+Flight.Distance, family = "binomial
", data = training)
summary(glm1)
```

```
##
## Call:
## glm(formula = satisfaction ~ Age + Gender + Customer.Type + Flight.Distance,
##     family = "binomial", data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4975  -0.9364  -0.5348   0.9401   2.1202
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -5.638e-01  2.726e-02 -20.683   <2e-16 ***
## Age                         -9.249e-05  4.534e-04  -0.204    0.838
## GenderMale                  -1.205e+00  1.371e-02 -87.908   <2e-16 ***
## Customer.TypeLoyal Customer  1.297e+00  1.812e-02  71.572   <2e-16 ***
## Flight.Distance             -6.180e-05  7.131e-06  -8.666   <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: 142659  on 103903  degrees of freedom
## Residual deviance: 128696  on 103899  degrees of freedom
## AIC: 128706
##
## Number of Fisher Scoring iterations: 4
```

Because p is roughly 0.67, we see that there is a 67% chance that this type of “person” will be satisfied with their trip given the certain attributes

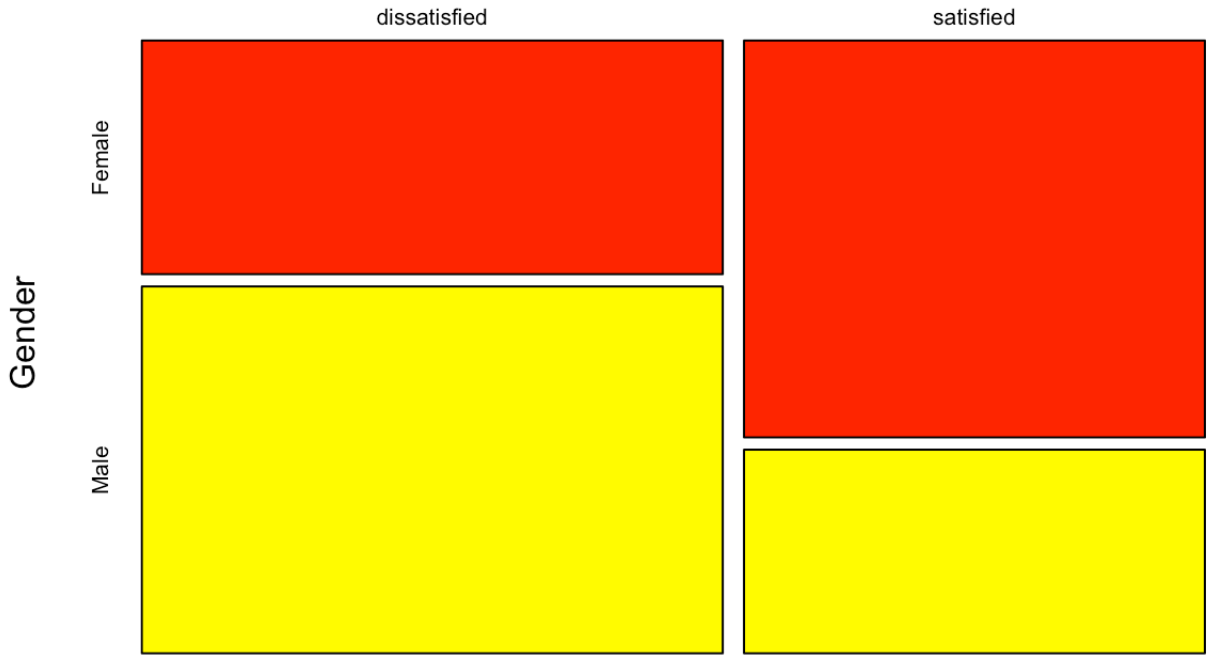
```
x <- data.frame(Gender = "Female", Age = 60, Customer.Type = "Loyal Customer", Flight
.Distance = 1000)
p <- predict(glm1, x)
p
```

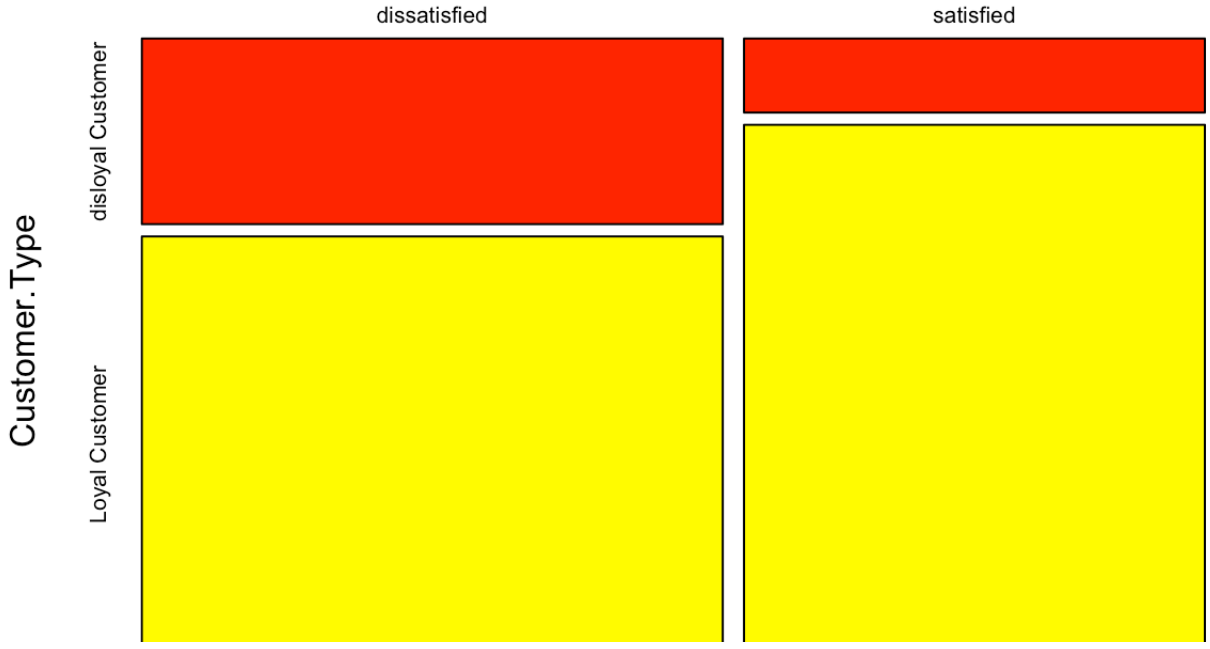
```
##        1
## 0.666102
```

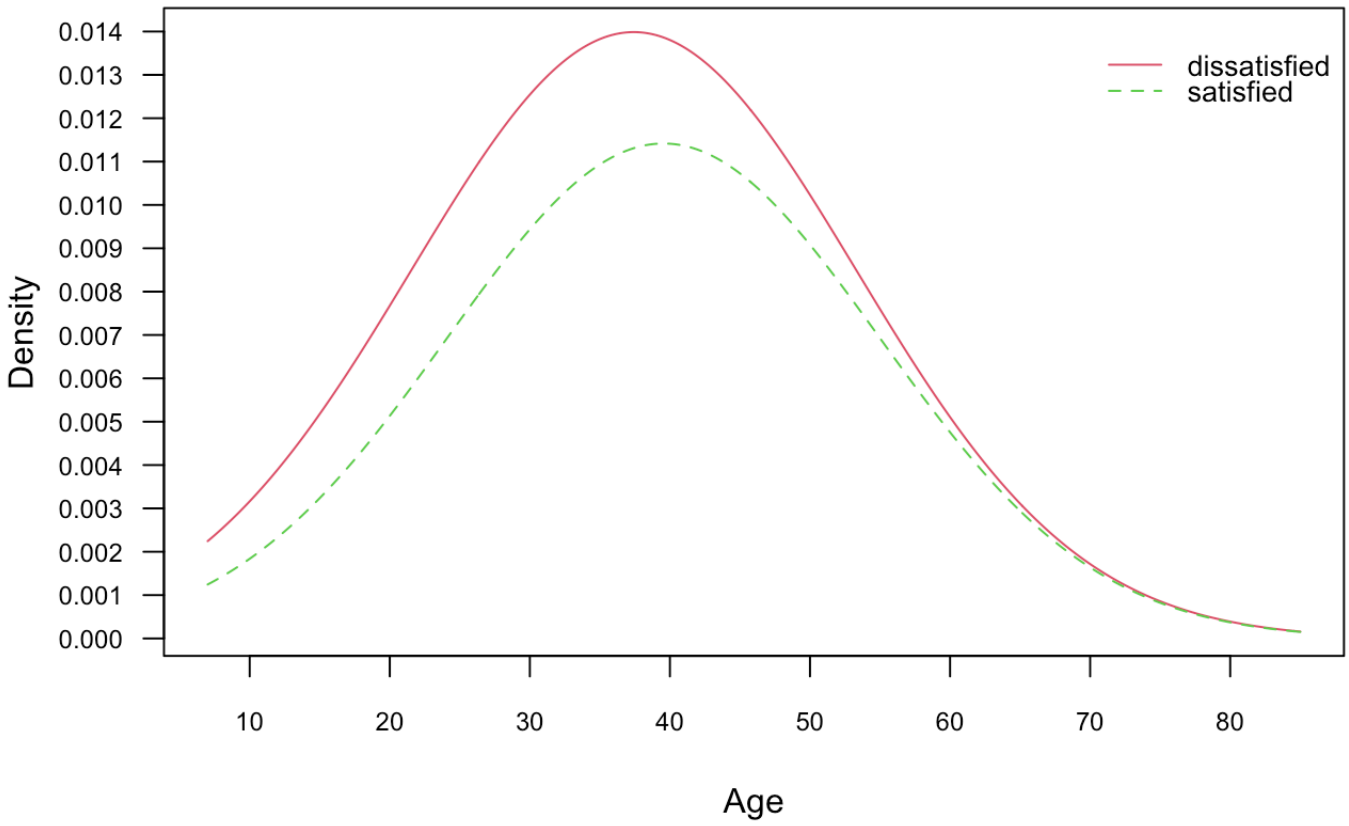
e. Build a naïve Bayes model and output what the model learned. Write a thorough explanation of the data.

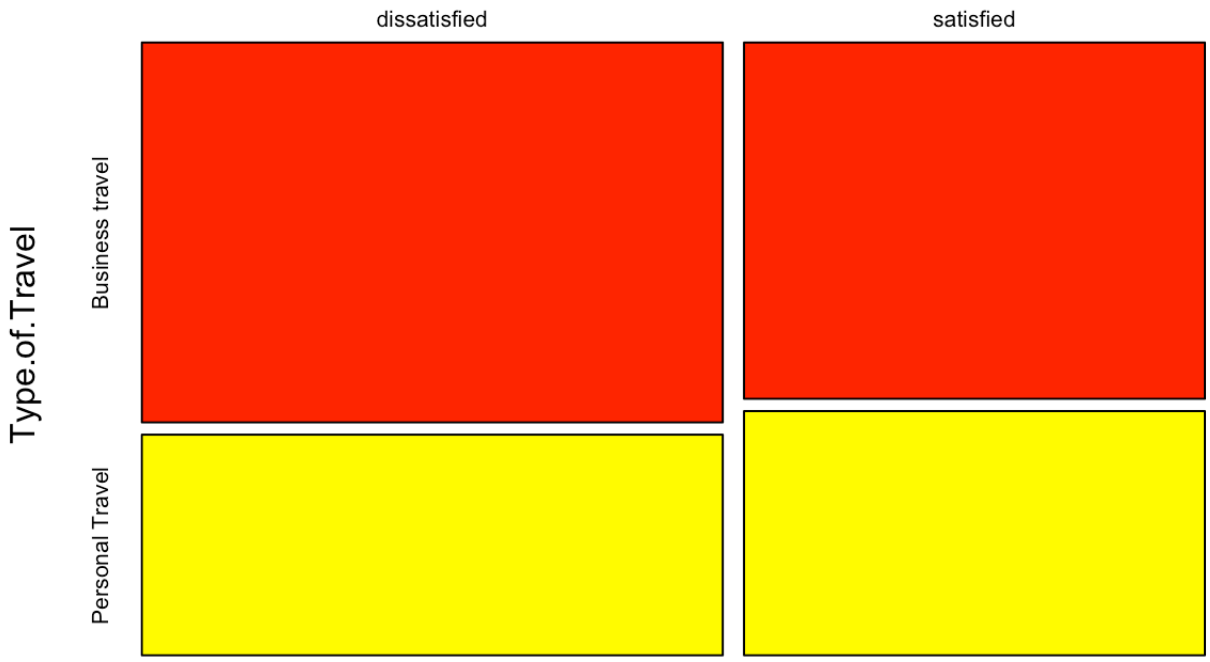
In this code snippet, the first line, will set the initial random value in order to make sure we get the same results for randomization and reproducability. In these next lines, we are creating the set we need in order to training the model and testing, and then create the naïve bayes model.

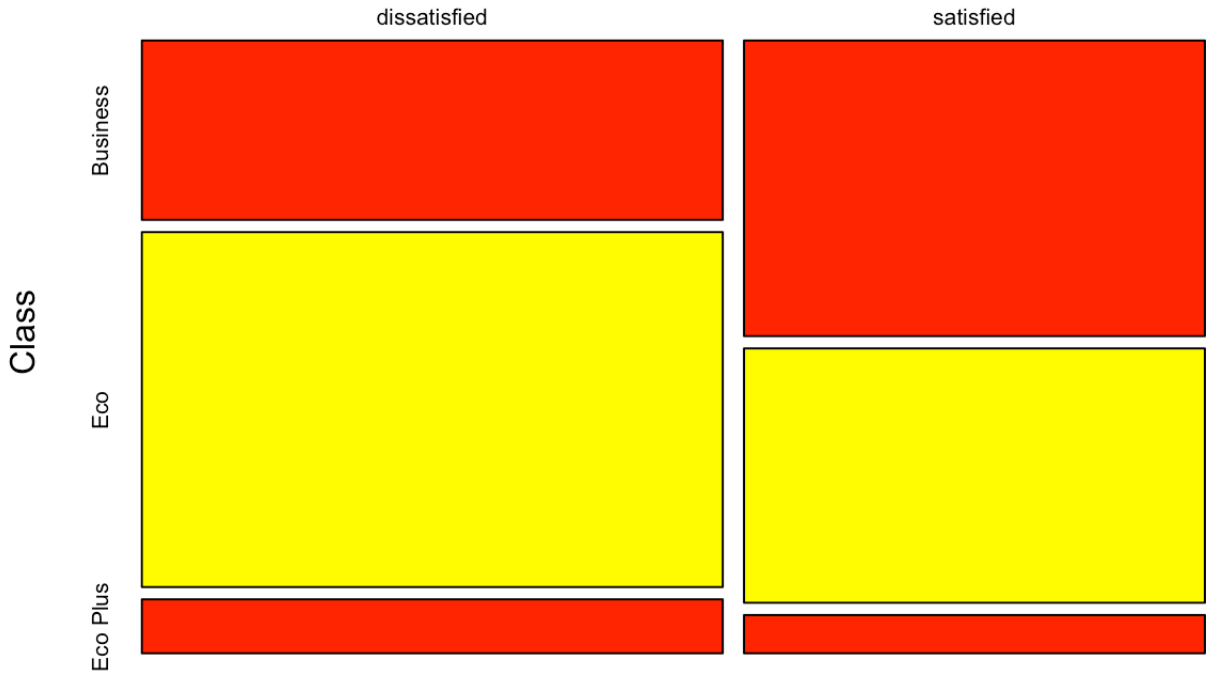
```
set.seed(1234)
ind <- sample(2, nrow(data), replace = T, prob = c(0.8, 0.2))
model <- naive_bayes(satisfaction~., data = training, usekernal = T)

plot(model)
```

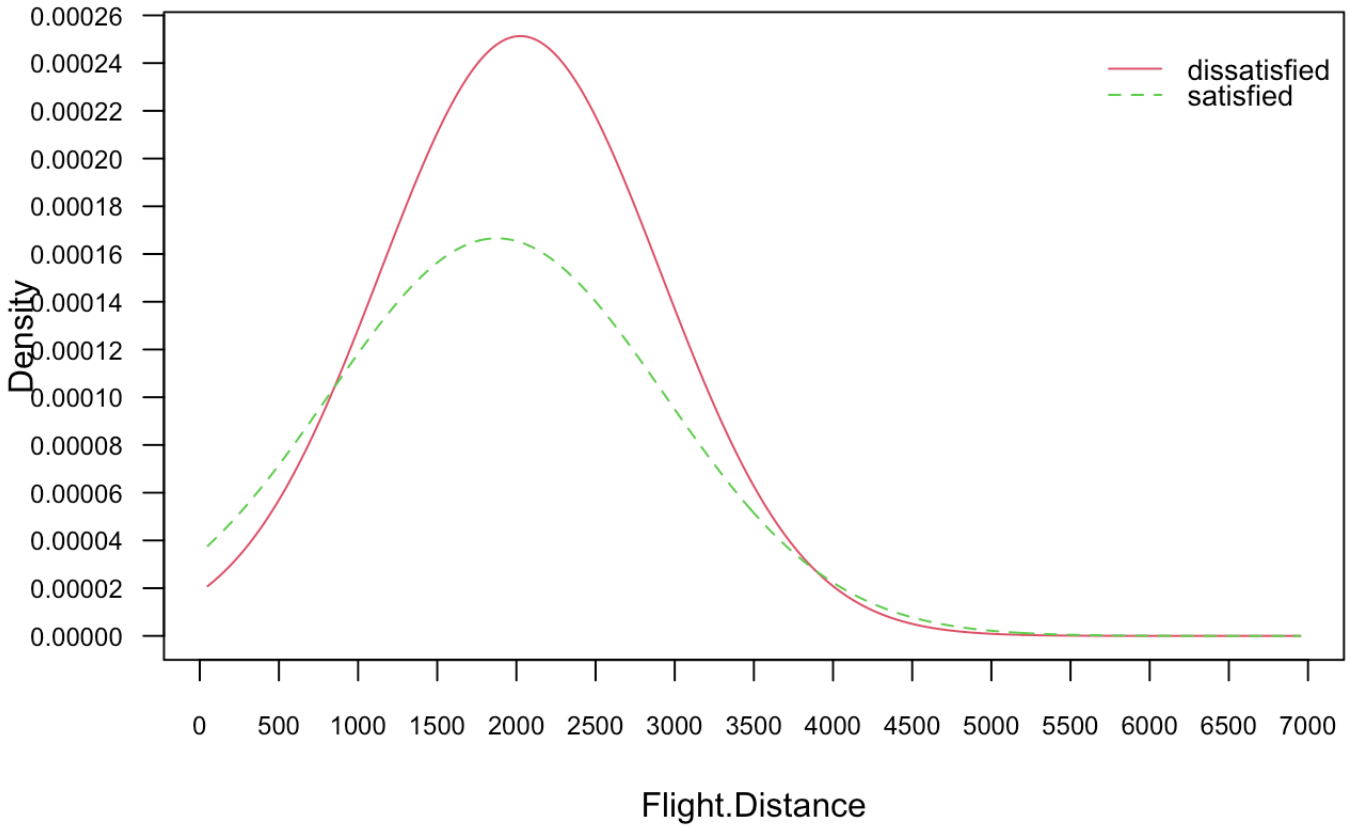

dissatisfied
satisfied
Gender
Female
Male

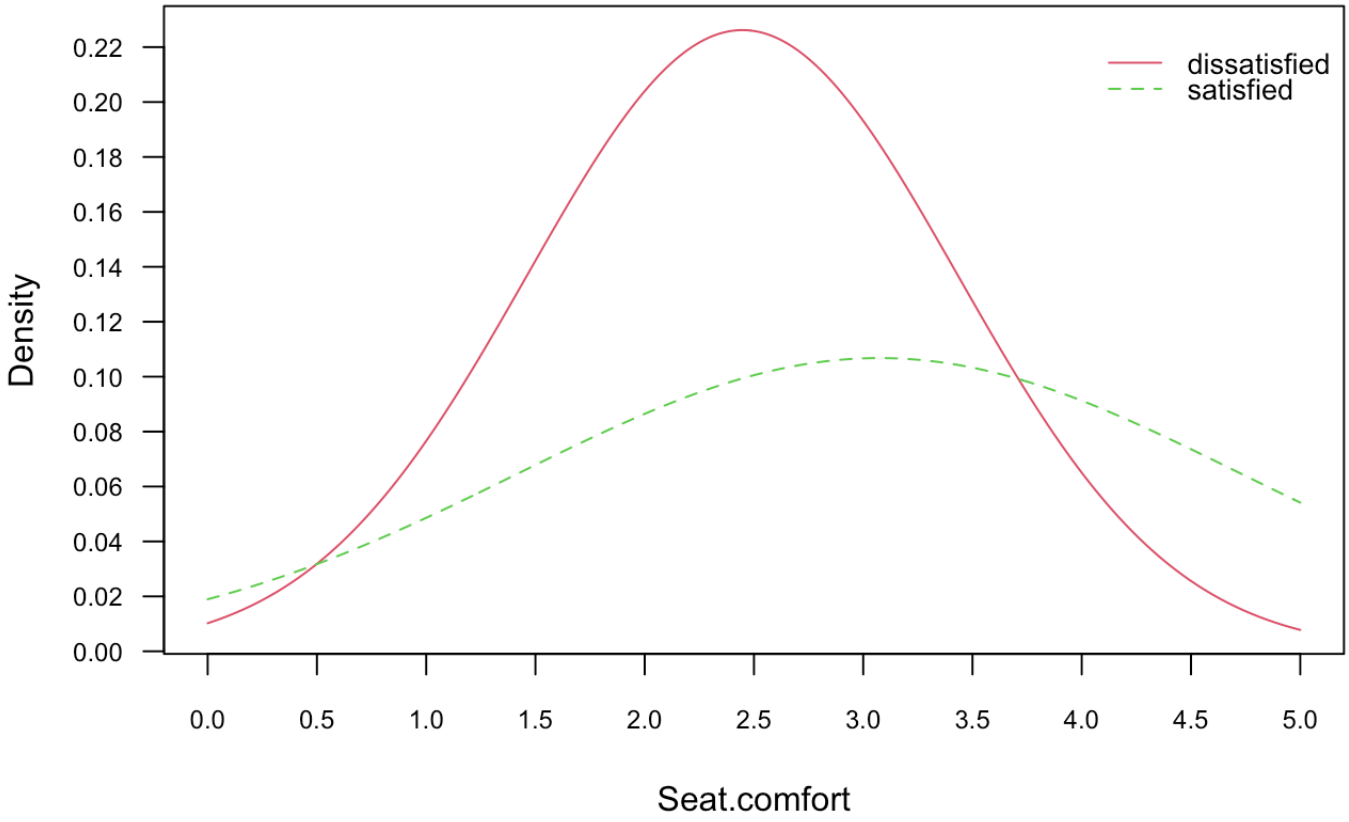

dissatisfied
satisfied
Customer.Type
disloyal Customer
Loyal Customer

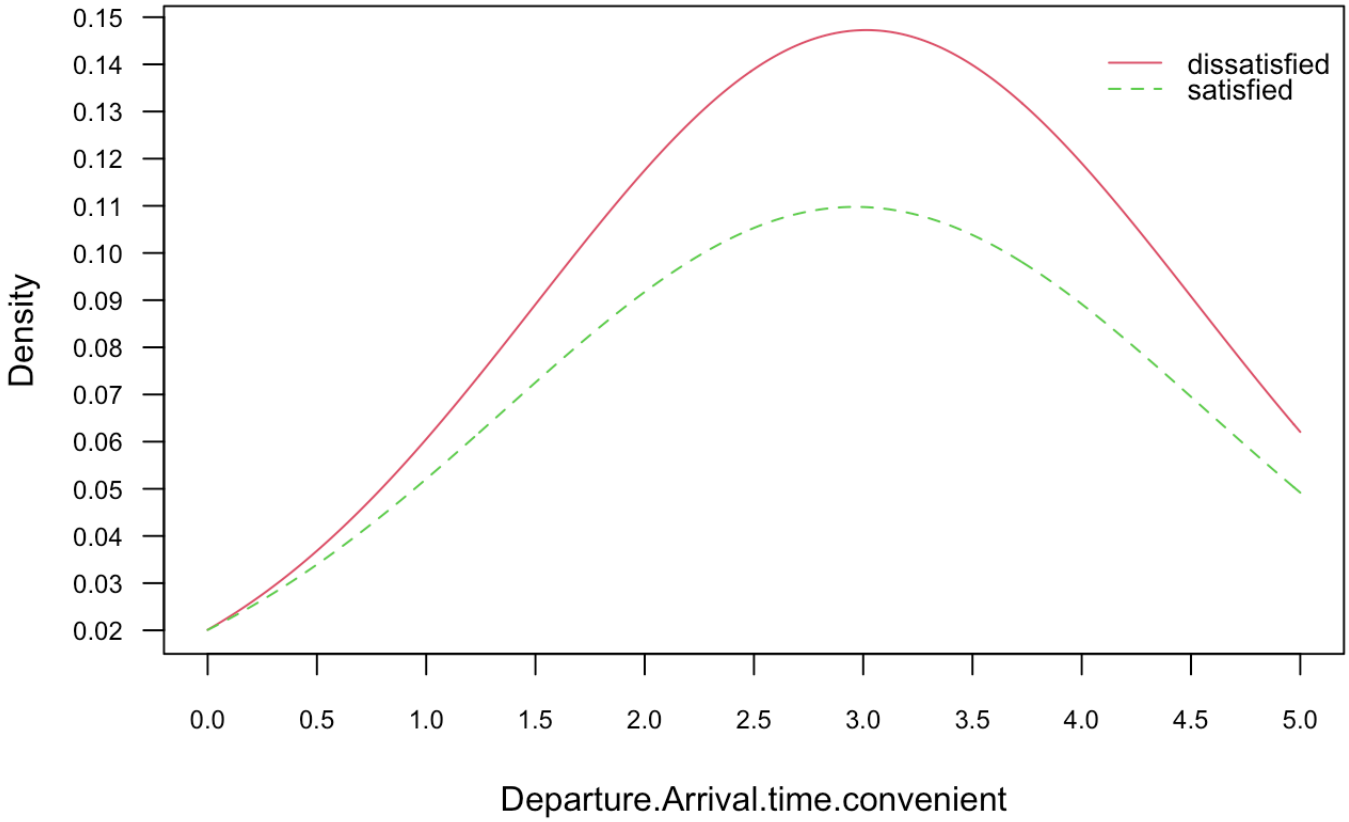

0.014
0.013
0.012
0.011
0.010
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Density
10
20
30
40
50
60
70
80
Age
dissatisfied
satisfied

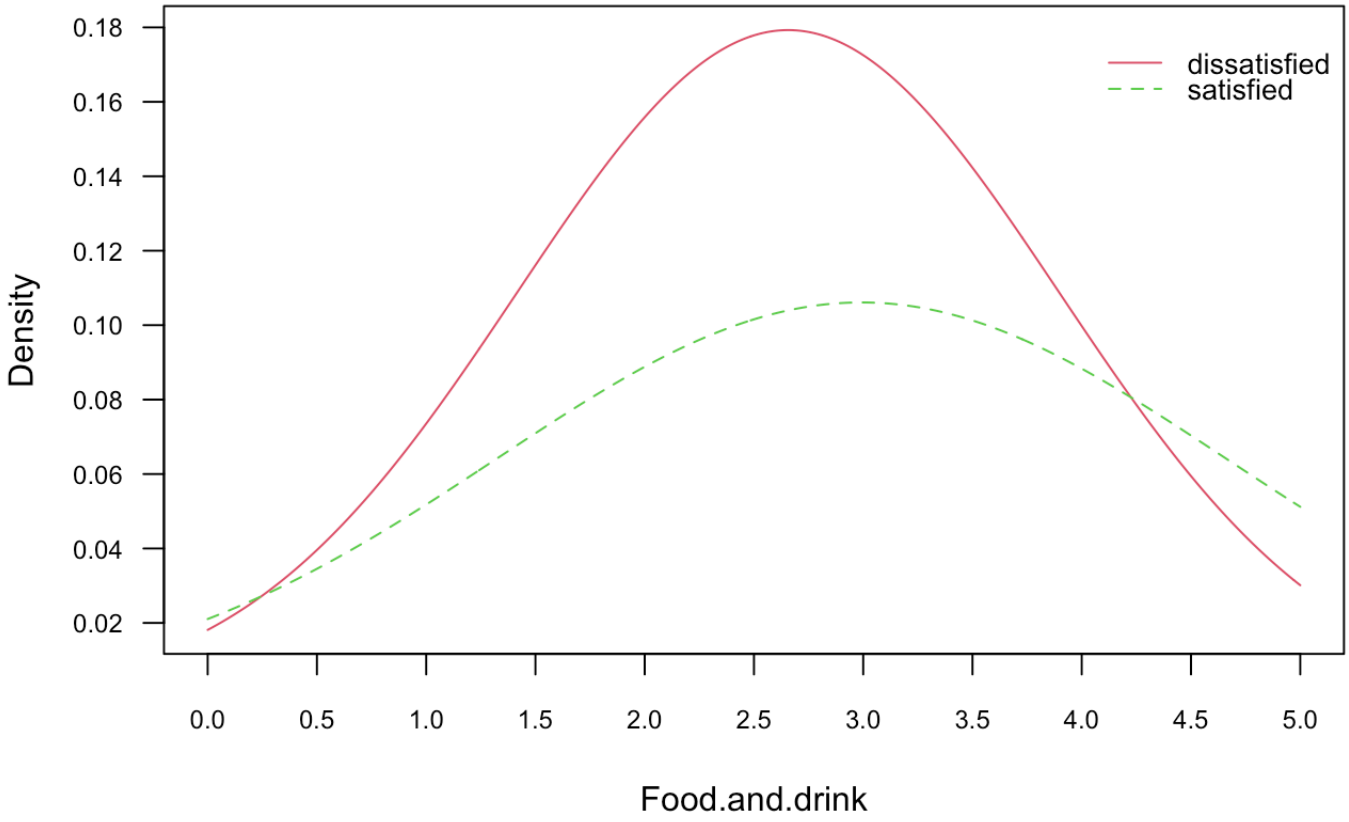

dissatisfied
satisfied
Type.of.Travel
Business travel
Personal Travel

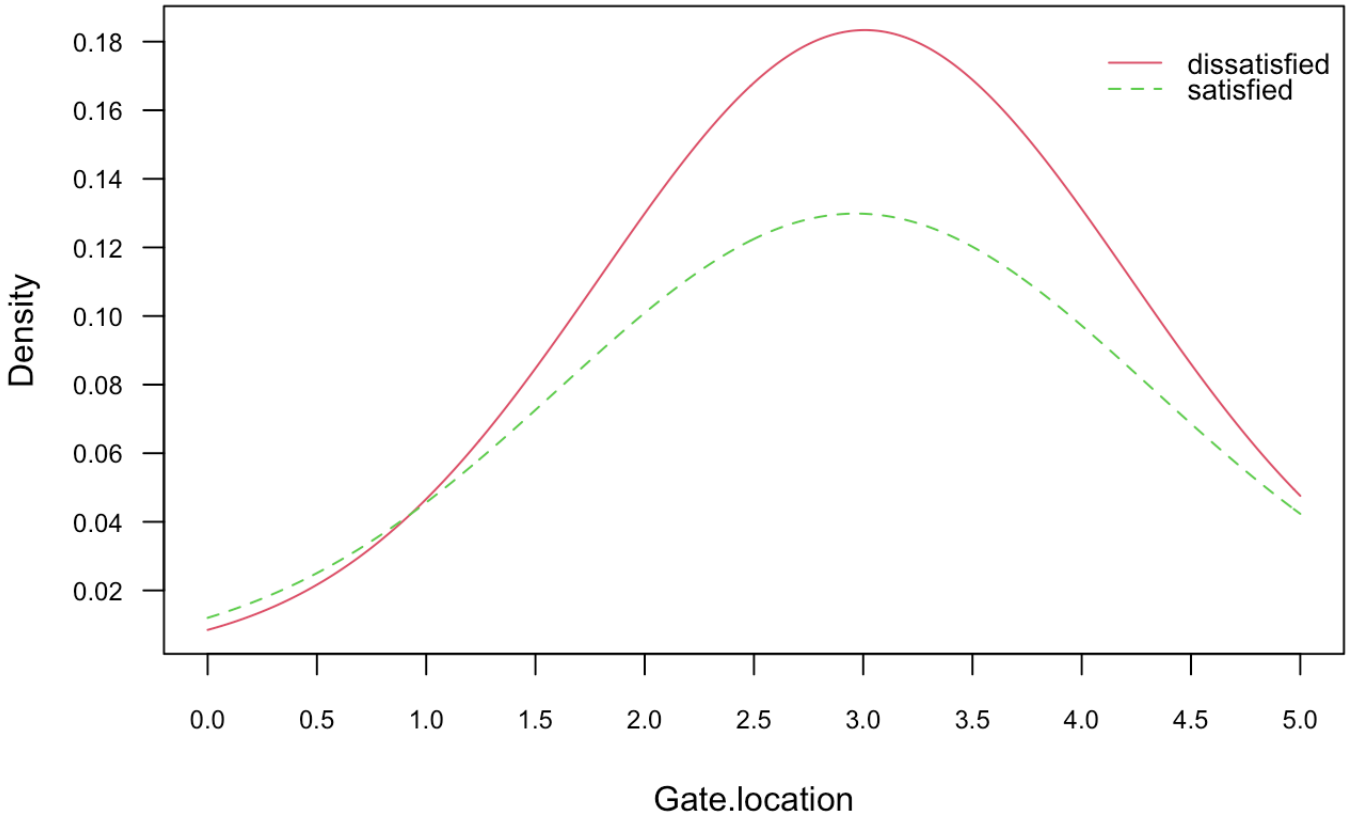

dissatisfied
satisfied
Class
Business
Eco
Eco Plus

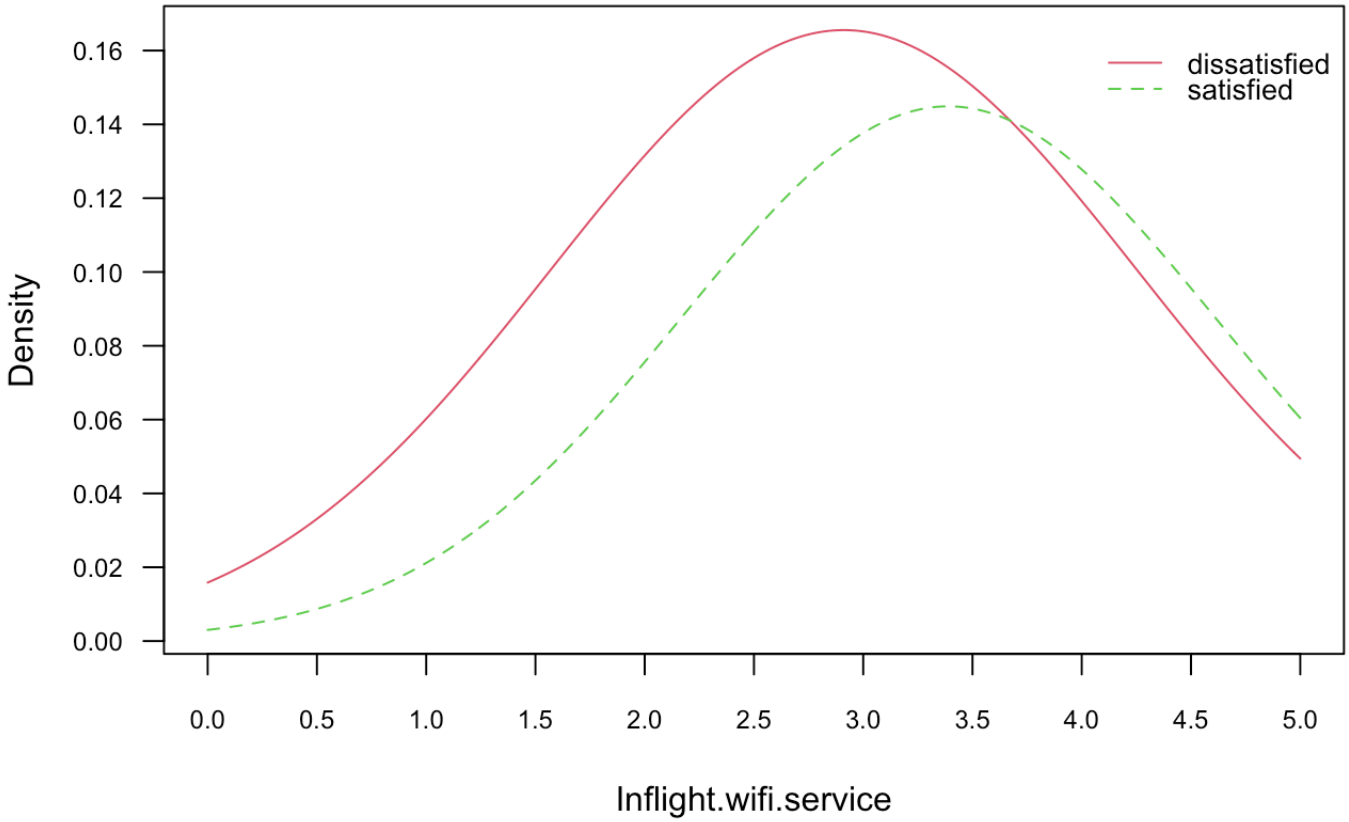

0.00026
0.00024
0.00022
0.00020
0.00018
0.00016
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
Density
dissatisfied
satisfied
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
Flight.Distance

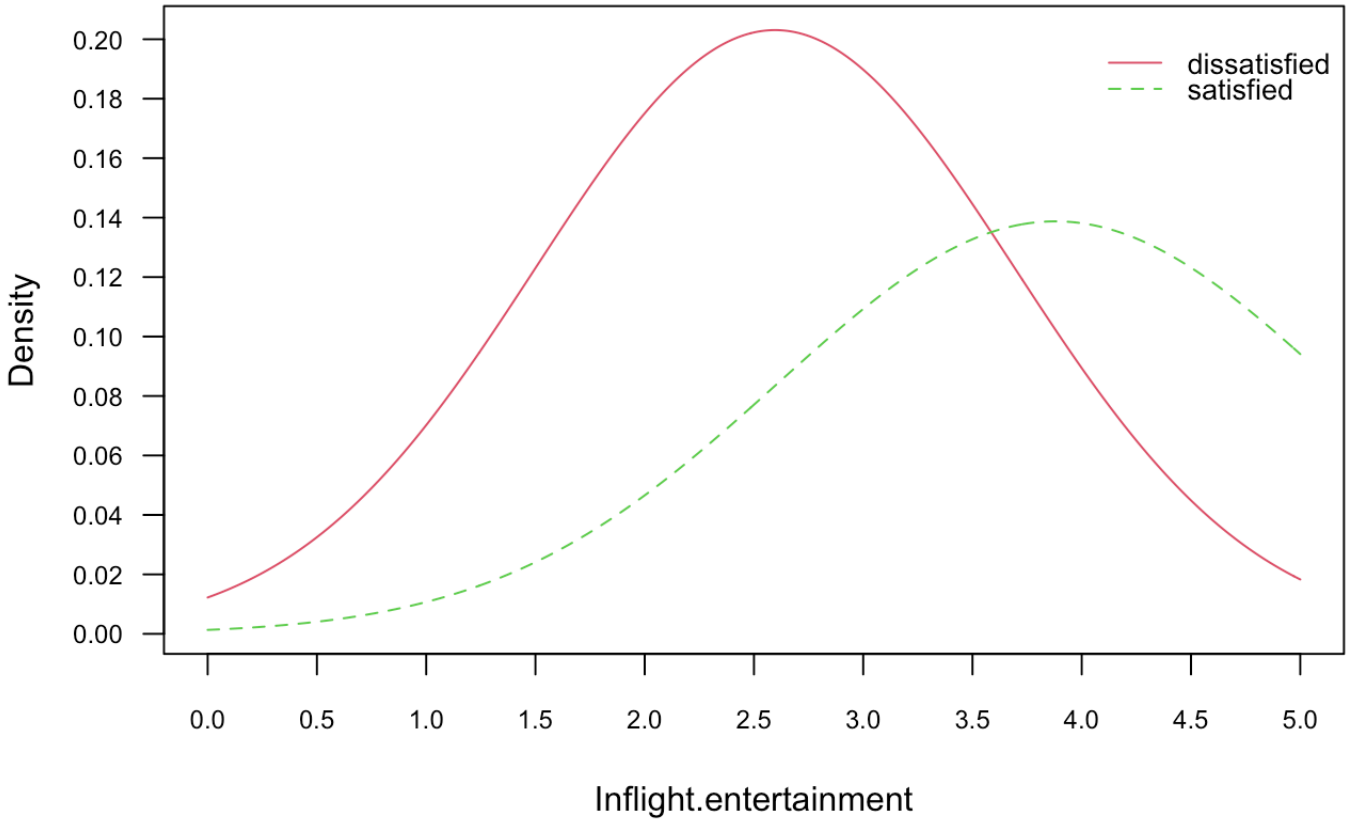

dissatisfied
satisfied
Density
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Seat.comfort

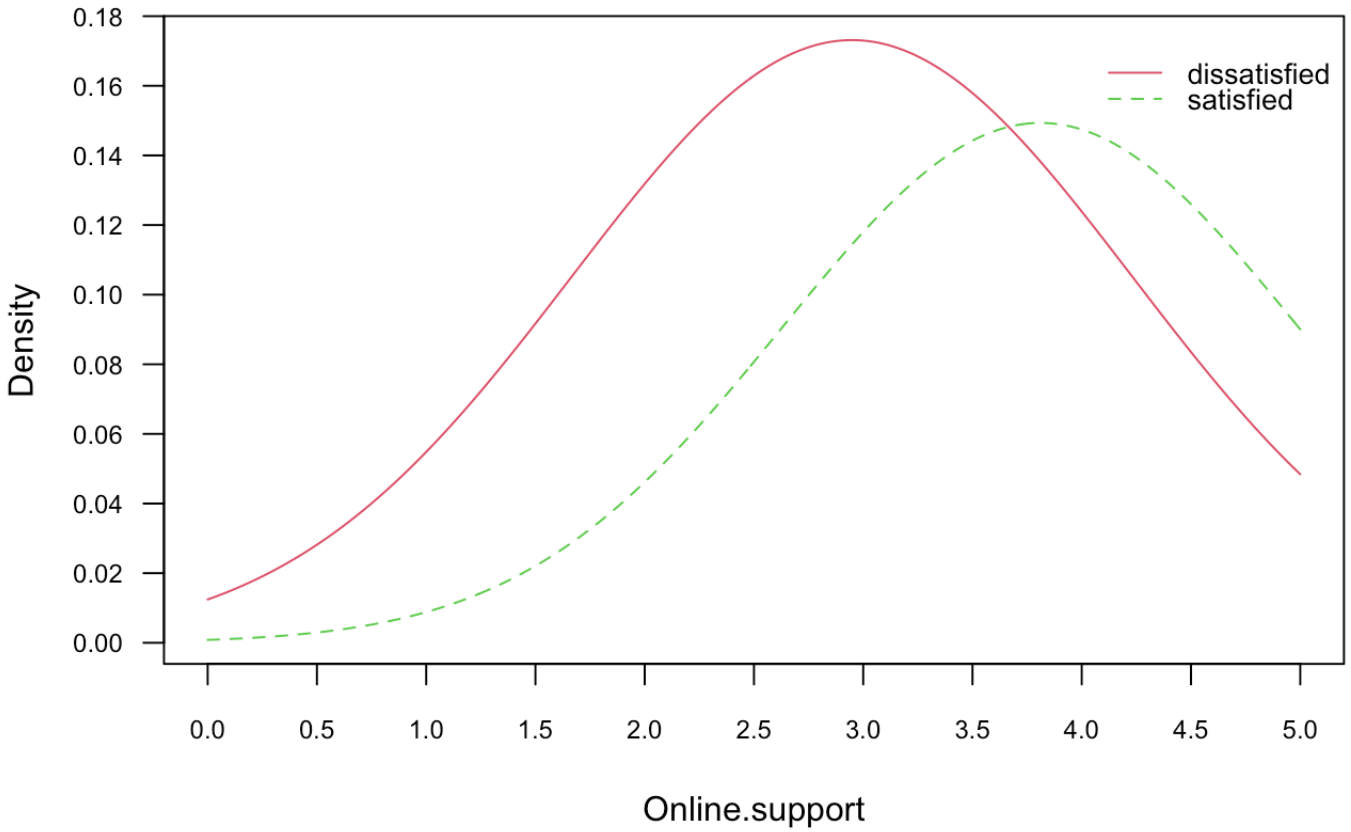

0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
Density
dissatisfied
satisfied
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Departure.Arrival.time.convenient

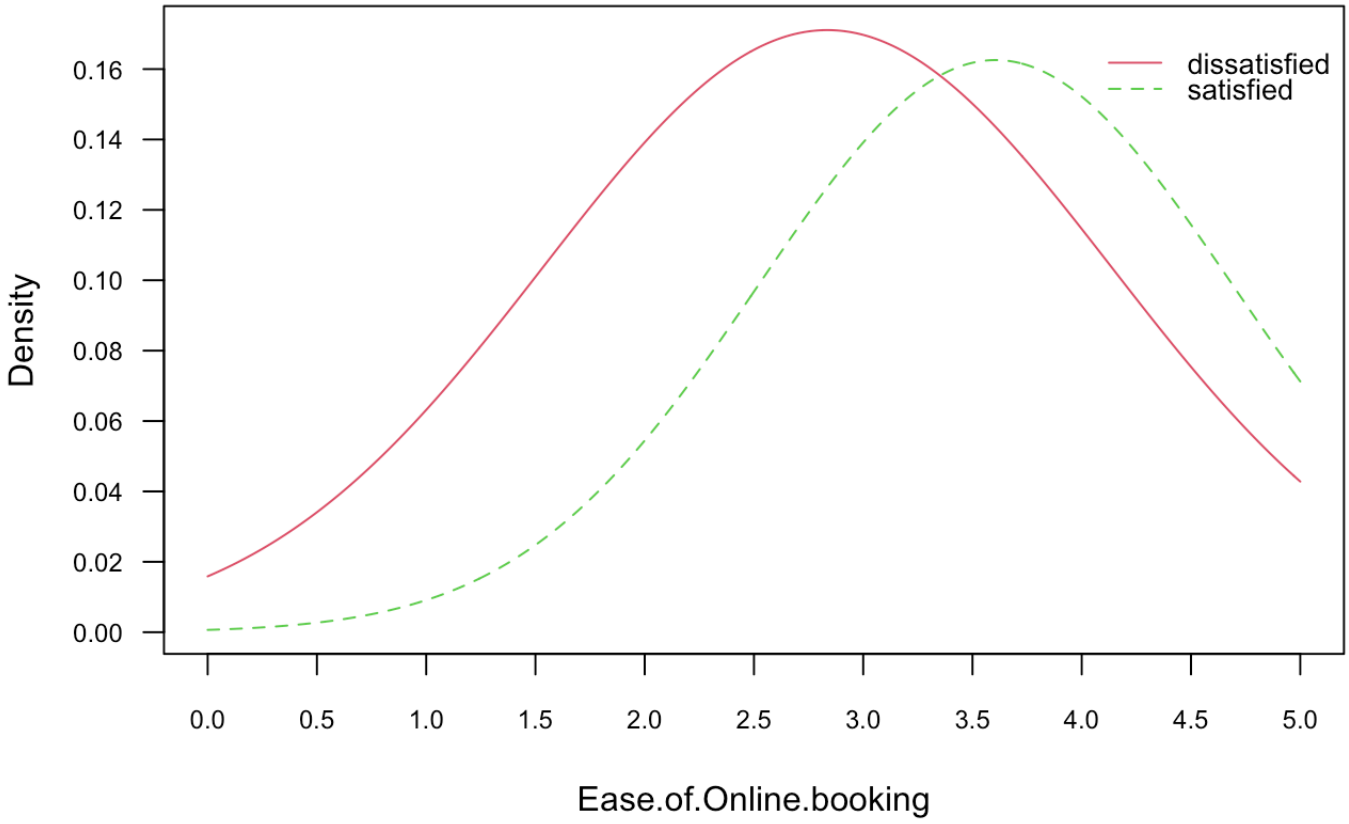

0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
Density
dissatisfied
satisfied
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Food.and.drink

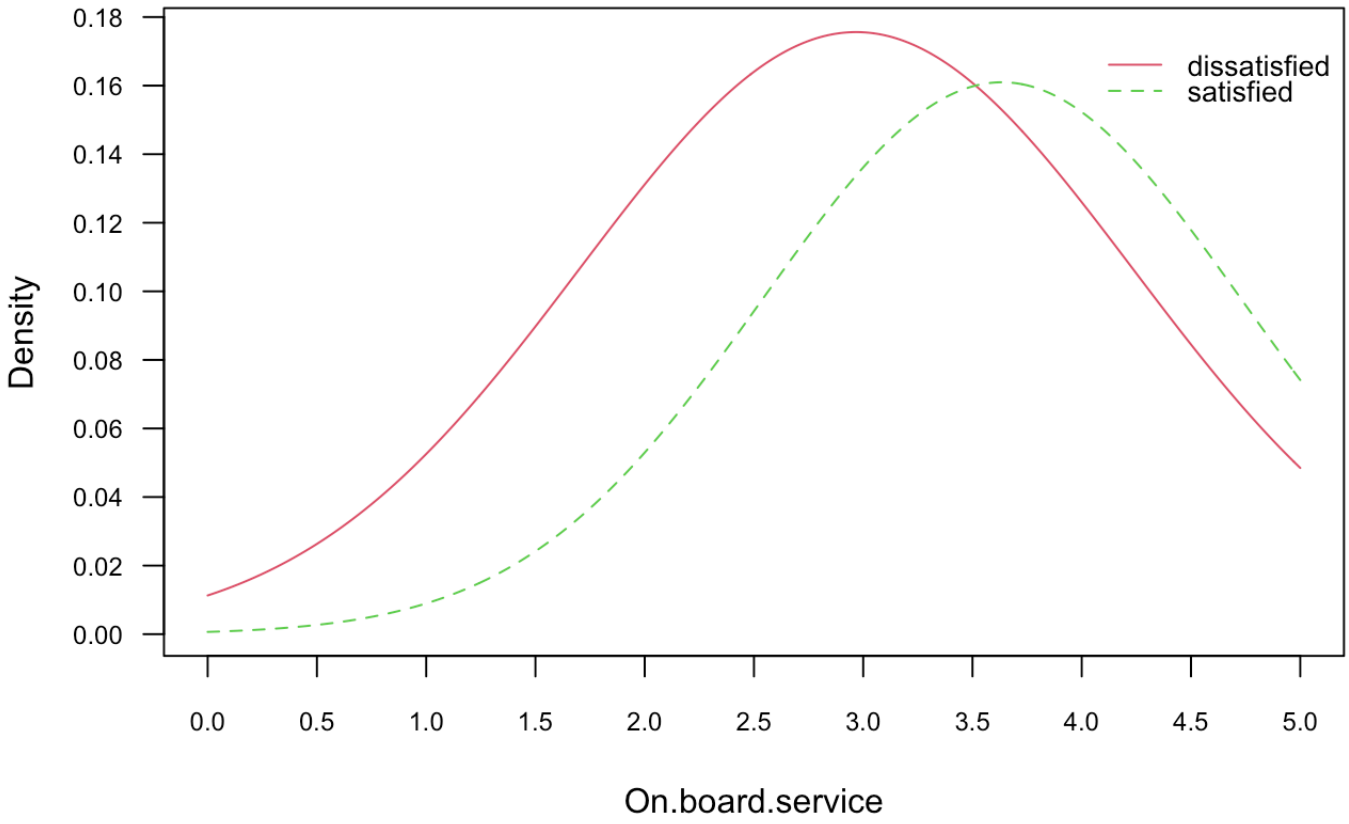

dissatisfied
satisfied
Density
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Gate.location

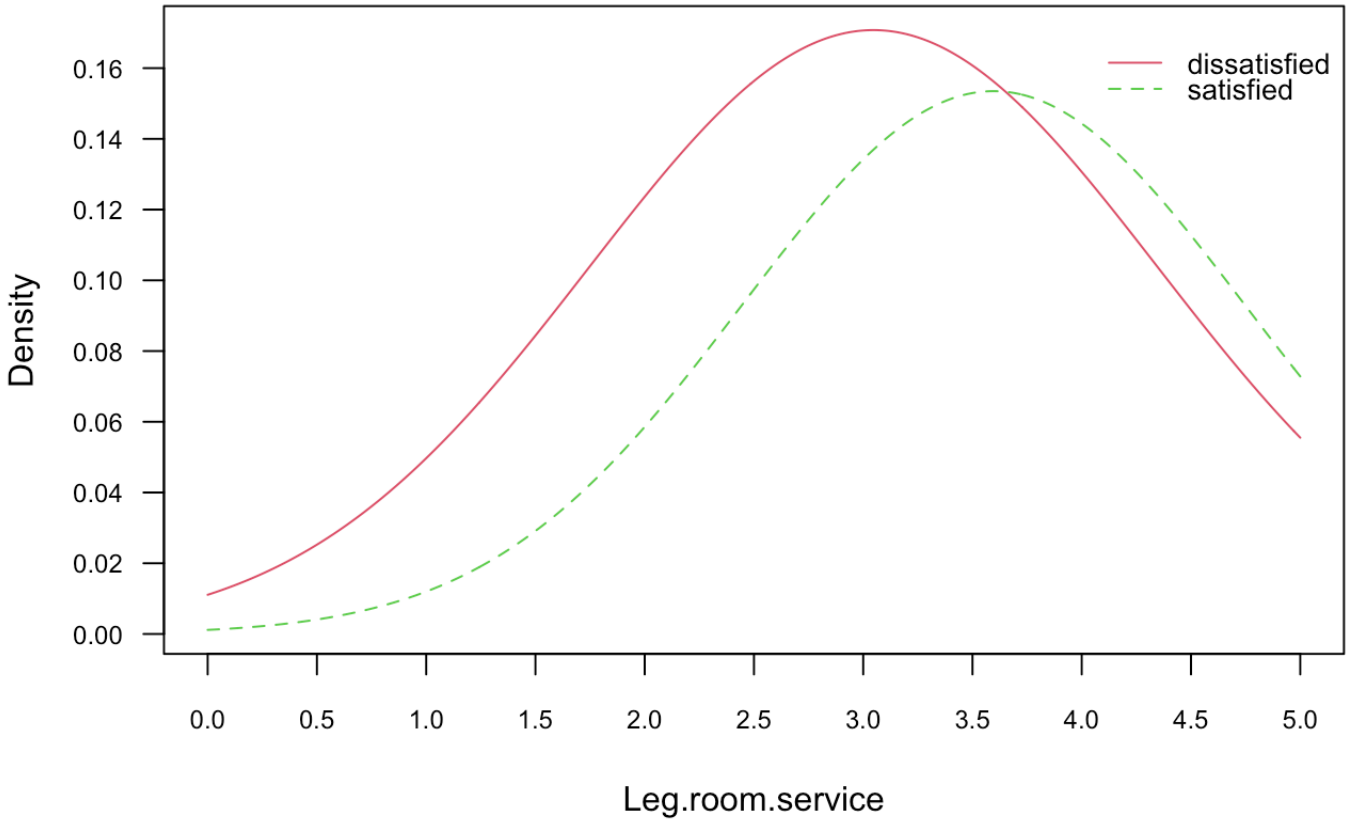

dissatisfied
satisfied
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Inflight.wifi.service

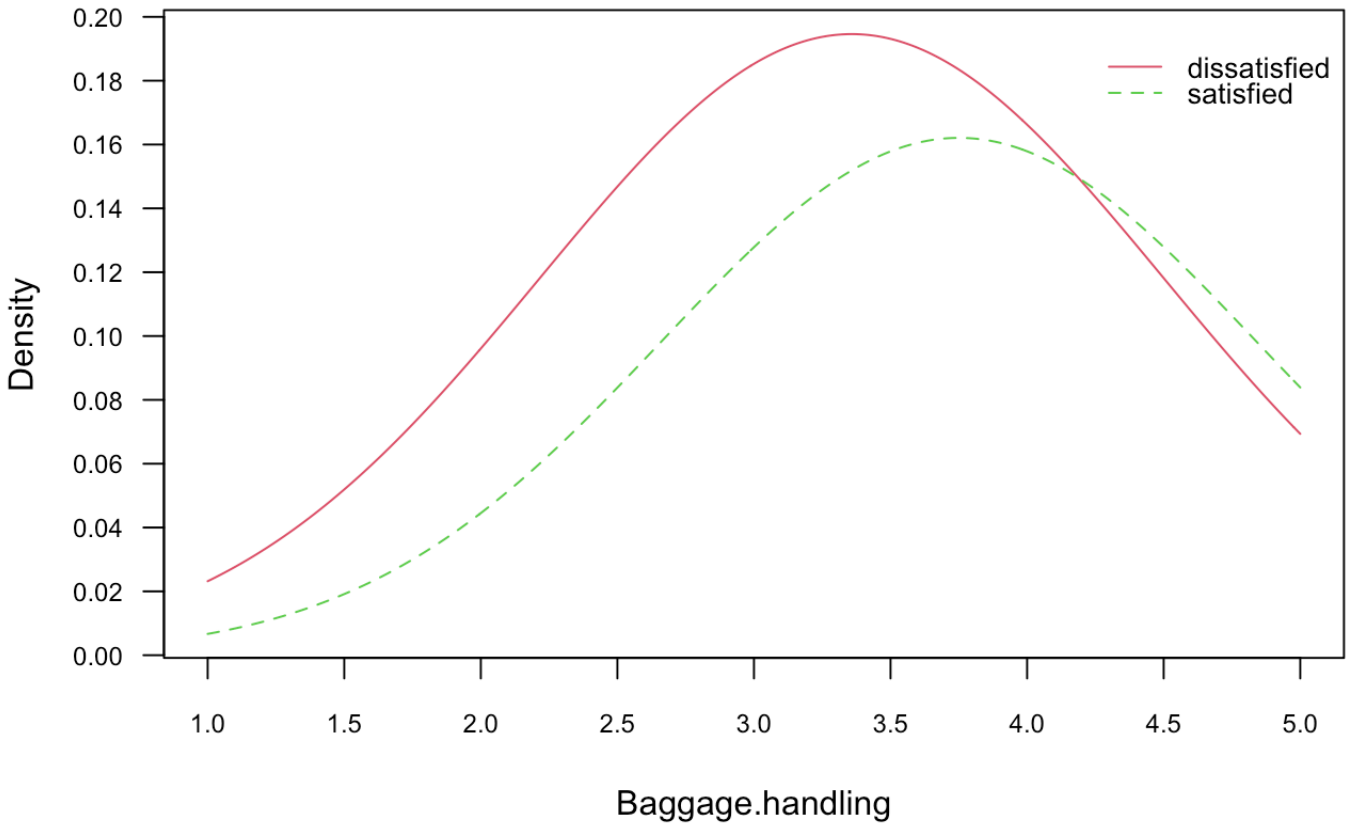

dissatisfied
satisfied
Density
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Inflight.entertainment

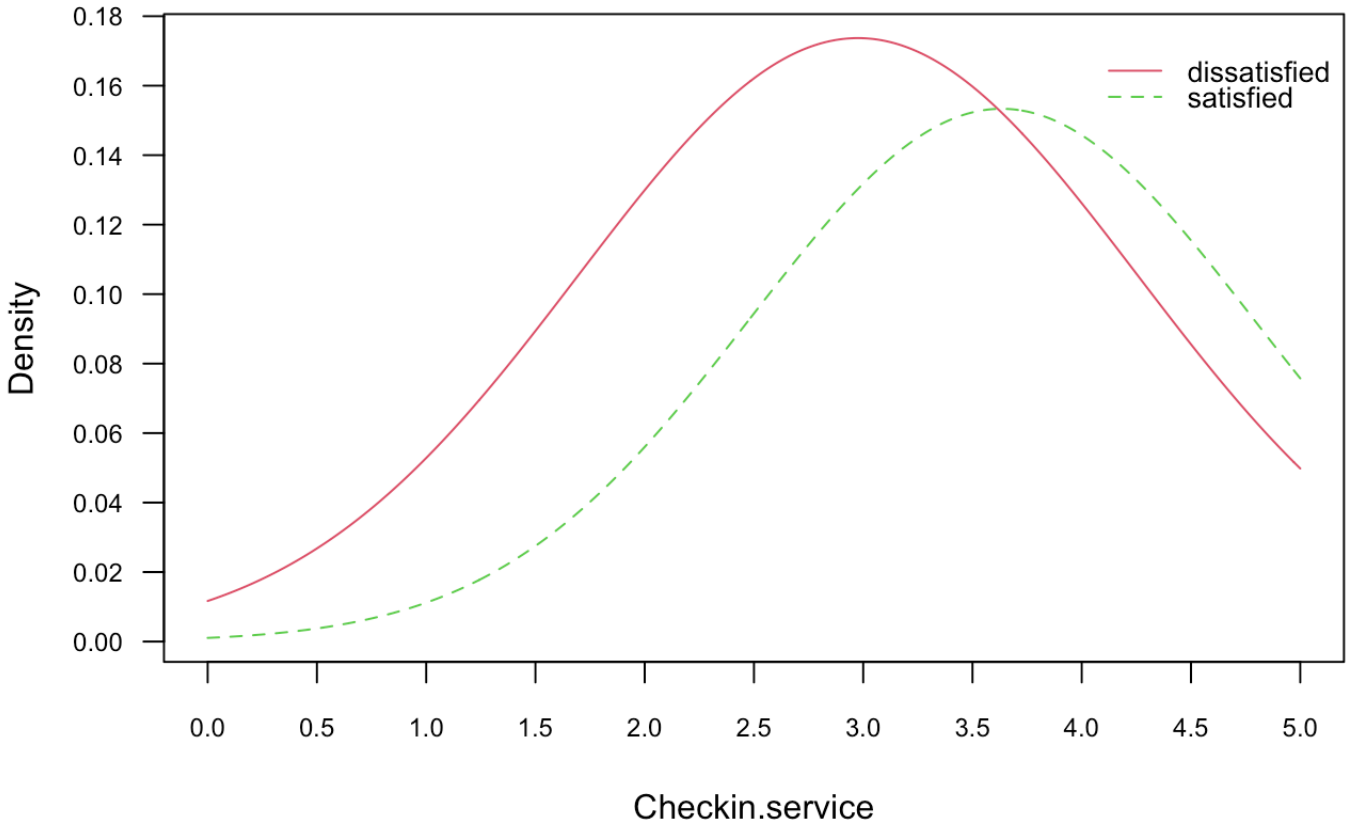

0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
dissatisfied
satisfied
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Online.support

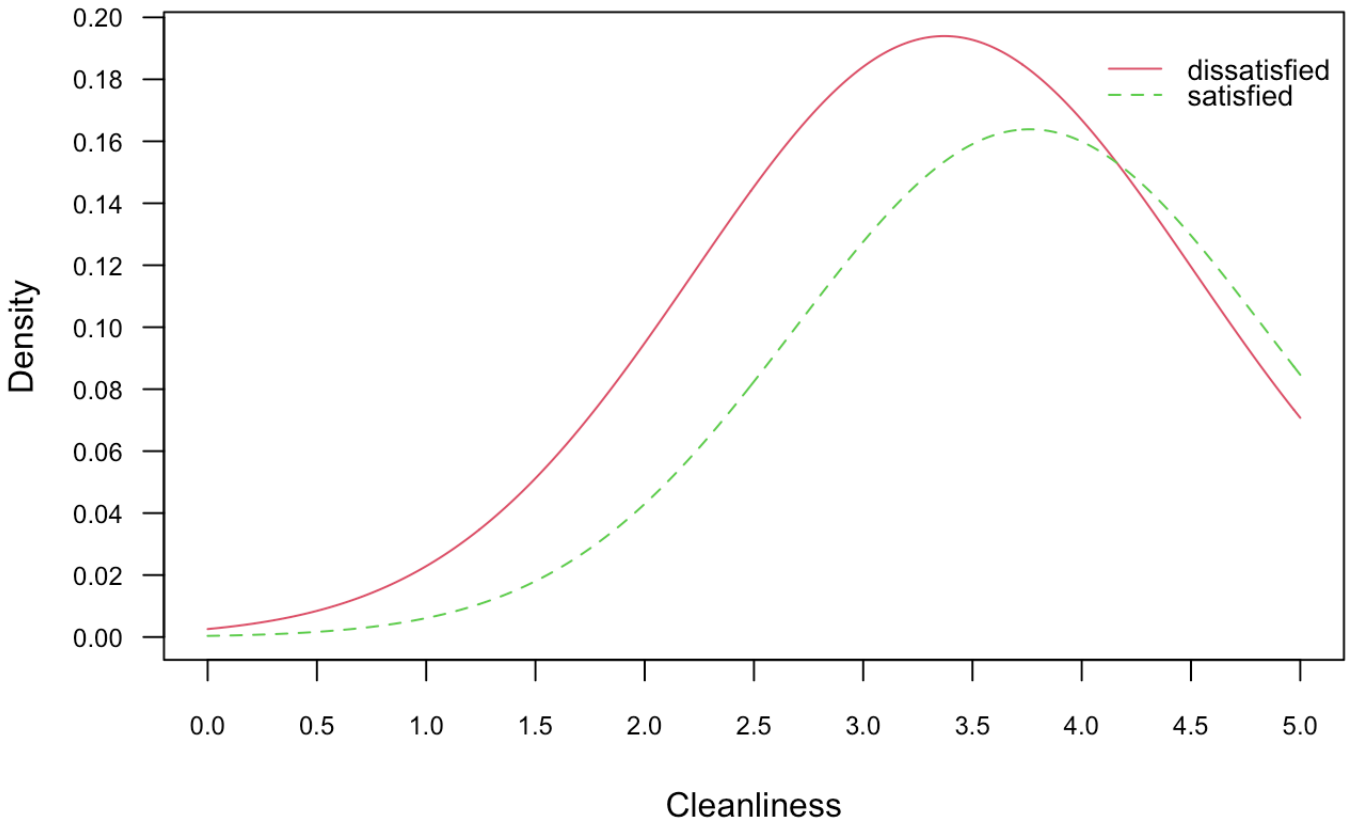

dissatisfied
satisfied
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Ease.of.Online.booking

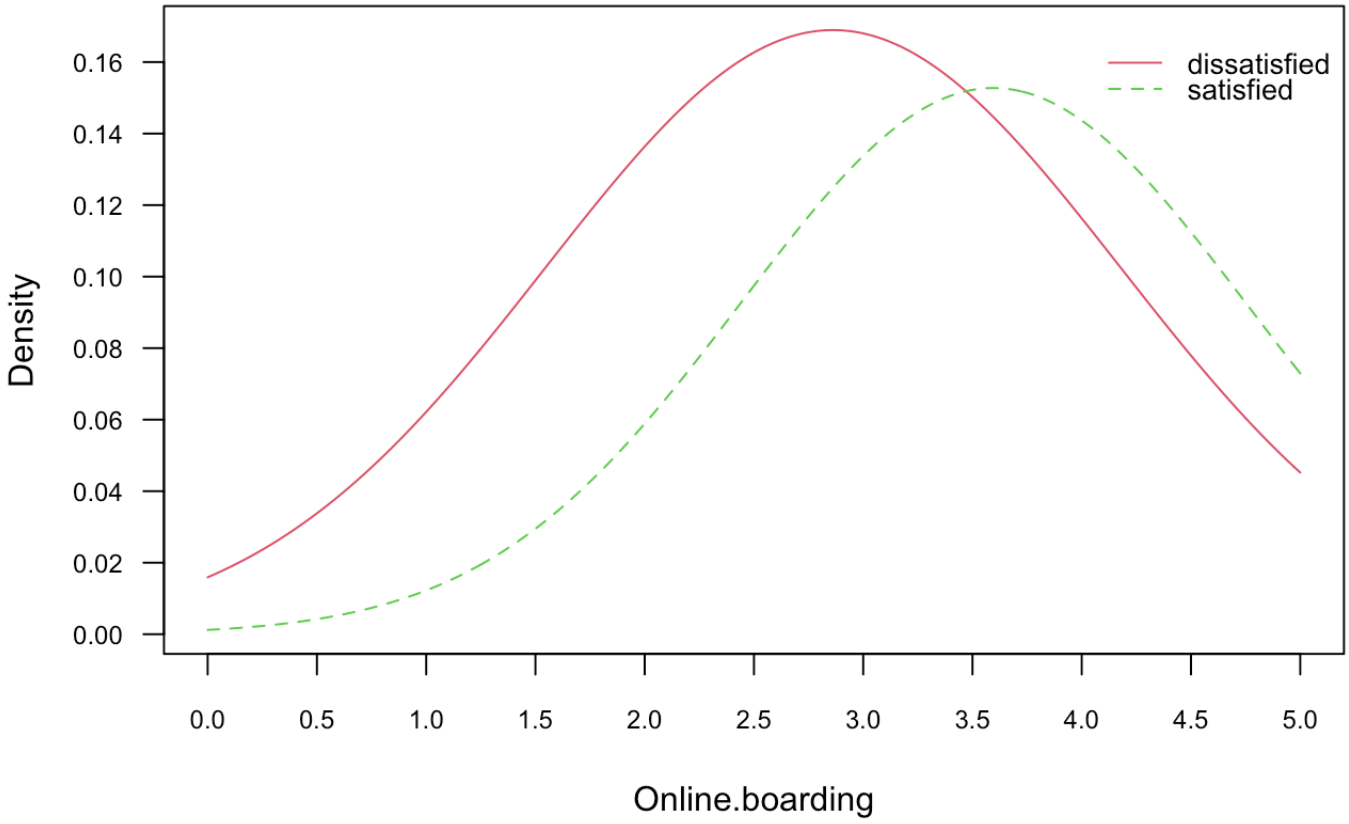

0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
dissatisfied
satisfied
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
On.board.service

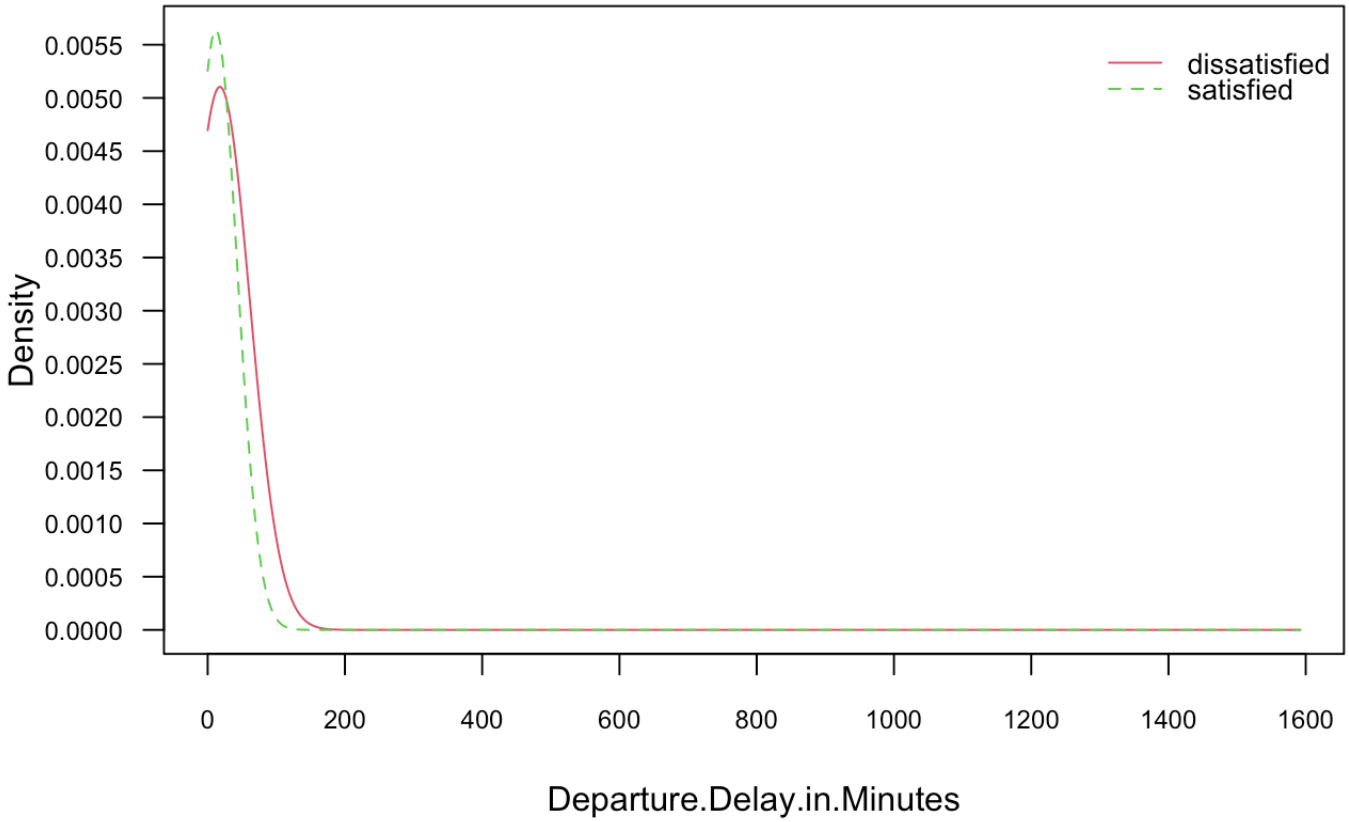

dissatisfied
satisfied
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Leg.room.service

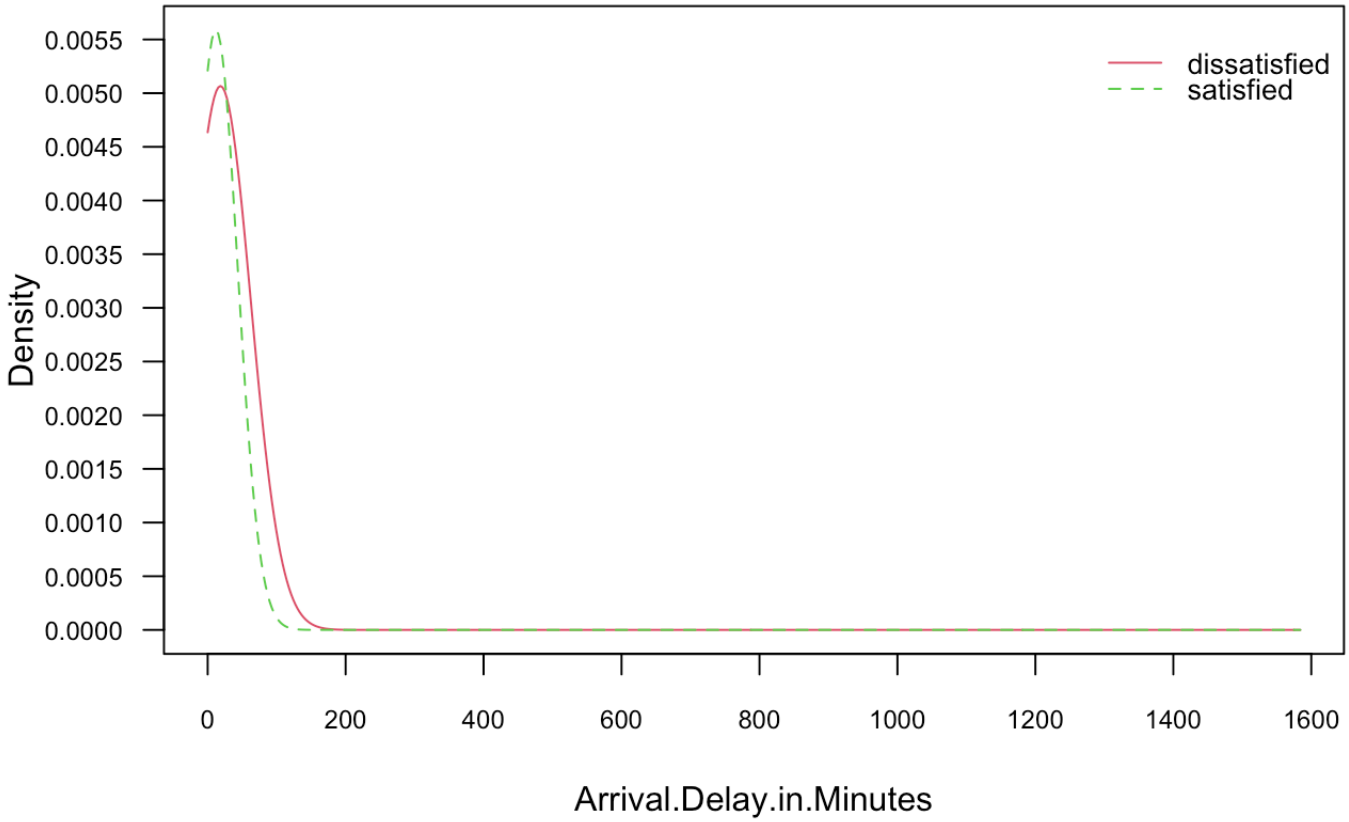

0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
dissatisfied
satisfied
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Baggage.handling


0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
dissatisfied
satisfied
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Checkin.service


0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Density
dissatisfied
satisfied
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Cleanliness


dissatisfied
satisfied
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Online.boarding


dissatisfied
satisfied
Density
0.0055
0.0050
0.0045
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
200
400
600
800
1000
1200
1400
1600
Departure.Delay.in.Minutes



f. Using these two classification models models, predict and evaluate on the test data using all of the classification metrics discussed in class. Compare the results and indicate why you think these results happened.

In the code here, we first load the e1071 library which contains the naiveBayes() function. This chart shows that the prior for Satisfaction 0.558 for being dissatisfied and 0.442 for being satisfied, with the likelihood data being shown as the conditional probabilities.

```
library(e1071)
nb1 <- naiveBayes(satisfaction~., data=training)
nb1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
```

```
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
## dissatisfied    satisfied
##    0.5576109    0.4423891
##
## Conditional probabilities:
##               Gender
## Y                 Female      Male
##   dissatisfied 0.3891056 0.6108944
##   satisfied    0.6609015 0.3390985
##
##               Customer.Type
## Y              disloyal Customer Loyal Customer
##   dissatisfied         0.3091408      0.6908592
##   satisfied            0.1233085      0.8766915
##
##               Age
## Y                  [,1]     [,2]
##   dissatisfied 37.42202 15.90451
##   satisfied    39.54218 15.45892
##
##               Type.of.Travel
## Y              Business travel Personal Travel
##   dissatisfied       0.6325900       0.3674100
##   satisfied          0.5933081       0.4066919
##
##               Class
## Y                Business        Eco   Eco Plus
##   dissatisfied 0.30486037 0.60336912 0.09177051
##   satisfied    0.50252360 0.43240656 0.06506983
##
##               Flight.Distance
## Y                  [,1]      [,2]
##   dissatisfied 2025.107  885.1598
##   satisfied    1876.354 1059.5364
##
##               Seat.comfort
## Y                  [,1]      [,2]
##   dissatisfied 2.447288 0.9834532
##   satisfied    3.073772 1.6529583
##
##               Departure.Arrival.time.convenient
## Y                  [,1]     [,2]
##   dissatisfied 3.014326 1.510594
##   satisfied    2.963125 1.607719
```

```
##
##               Food.and.drink
## Y                  [,1]     [,2]
##   dissatisfied 2.656512 1.240858
##   satisfied    2.991994 1.663390
##
##               Gate.location
## Y                  [,1]     [,2]
##   dissatisfied 3.007404 1.213179
##   satisfied    2.965300 1.358982
##
##               Inflight.wifi.service
## Y                  [,1]     [,2]
##   dissatisfied 2.911043 1.343755
##   satisfied    3.389353 1.218130
##
##               Inflight.entertainment
## Y                  [,1]     [,2]
##   dissatisfied 2.596690 1.095417
##   satisfied    3.879193 1.272013
##
##               Online.support
## Y                  [,1]     [,2]
##   dissatisfied 2.948669 1.284908
##   satisfied    3.811056 1.181846
##
##               Ease.of.Online.booking
## Y                  [,1]     [,2]
##   dissatisfied 2.835617 1.300298
##   satisfied    3.605643 1.085655
##
##               On.board.service
## Y                  [,1]     [,2]
##   dissatisfied 2.967741 1.266721
##   satisfied    3.634665 1.096338
##
##               Leg.room.service
## Y                  [,1]     [,2]
##   dissatisfied 3.046878 1.302740
##   satisfied    3.596115 1.149793
##
##               Baggage.handling
## Y                  [,1]     [,2]
##   dissatisfied 3.358055 1.143116
##   satisfied    3.750294 1.088850
##
##               Checkin.service
```

```
## Y                   [,1]     [,2]
##   dissatisfied 2.975871 1.280703
##   satisfied    3.633229 1.150916
##
##               Cleanliness
## Y                   [,1]     [,2]
##   dissatisfied 3.371311 1.146929
##   satisfied    3.762063 1.076965
##
##               Online.boarding
## Y                   [,1]     [,2]
##   dissatisfied 2.861887 1.316736
##   satisfied    3.595353 1.155444
##
##               Departure.Delay.in.Minutes
## Y                  [,1]     [,2]
##   dissatisfied 17.8075 43.56644
##   satisfied    11.7484 31.29689
##
##               Arrival.Delay.in.Minutes
## Y                   [,1]     [,2]
##   dissatisfied 18.51160 43.90781
##   satisfied    11.81787 31.58097
```

Here are the raw probabilities, which in here, are more accurate than the logistic regression values

```
p2_raw <- predict(nb1, newdata=test, type="raw")
head(p2_raw, n=2)
```

```
##      dissatisfied satisfied
## [1,] 0.0053950434 0.9946050
## [2,] 0.0003631853 0.9996368
```

g. Write a paragraph listing the strengths and weaknesses of Naïve Bayes and Logistic Regression

Logistic Regression

- Advantages - Logistic Regression is easier to implement, however, if there are less observations than features, then it could lead to over-fitting of the graph. It also has relatively good accuracy for simple data sets and is pretty quick at classifying them.
- Disadvantage - One major disadvantage is logistic regression's assumption of a linear relationship between in the dependent and independent variables

Naïve Bayes

- Advantages - Quick and saves time, if the function assumes the independent variables hold true, then it can be more accurate with less data
- Disadvantages - This model assumes that all variables are independent, which is not always the case in a real life scenario, which means that the probability might not be as accurate as it may seem.

h. Write a paragraph listing the benefits, drawbacks of each of the classification metrics used, and briefly describe what each metric tells you.

Accuracy

The percentages of observations that were classified correctly. Although on a small and unbalanced data set it may seem accurate the accuracy might be skewed.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.