# Final Exam - Fundamentals of Machine Learning

Julia Thacker

12/01/2021

```
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(e1071)
airlinedata<-read.csv("airline_passenger_satisfaction.csv")
airlinedata<-na.omit(airlinedata)
```

Read the data and removed any NAs

```
airlinedata$Gender<-as.factor(airlinedata$Gender)
airlinedata$type_of_travel<-as.factor(airlinedata$type_of_travel)
airlinedata$customer_class<-as.factor(airlinedata$customer_class)
airlinedata$customer_type<-as.factor(airlinedata$customer_type)
airlinedata$satisfaction<-as.factor(airlinedata$satisfaction)
airlinedata$inflight_wifi_service<-
as.factor(airlinedata$inflight_wifi_service)
airlinedata$inflight_entertainment<-
as.factor(airlinedata$inflight_entertainment)
```

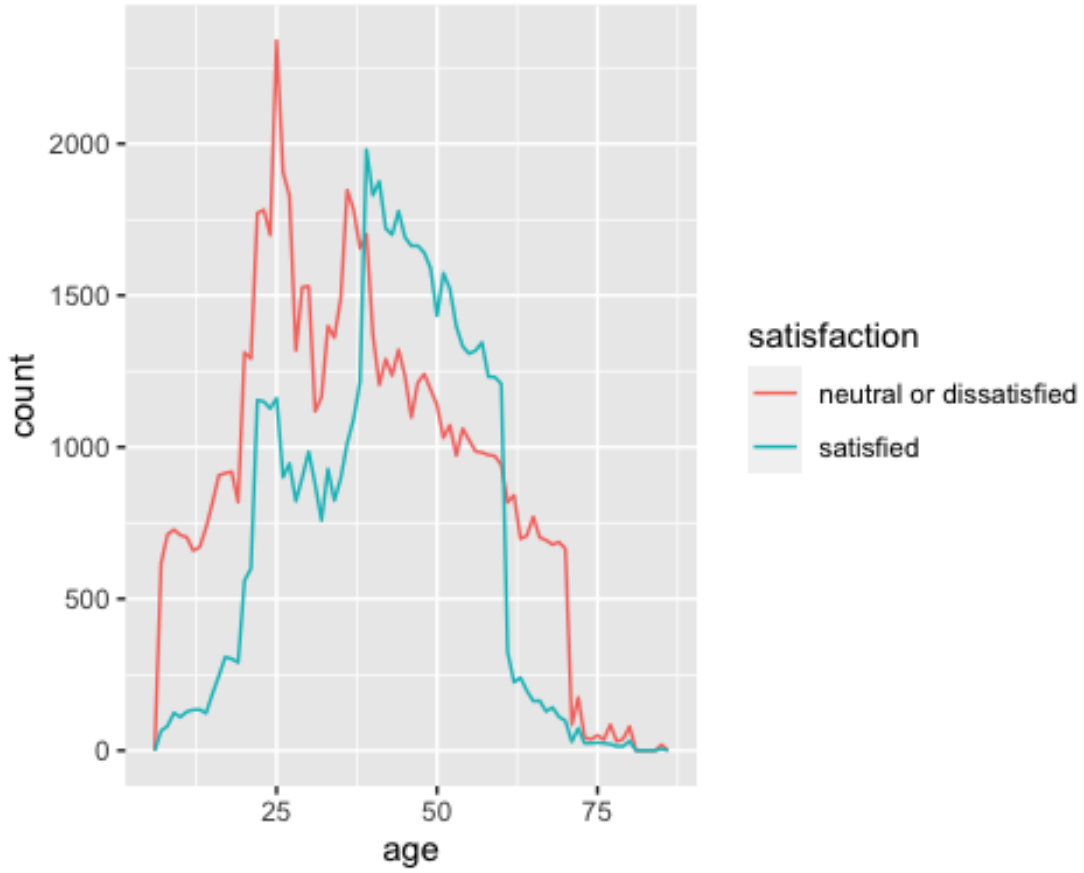
Converted necessary variables into factors

```
 str(airlinedata)

## 'data.frame':    129487 obs. of  24 variables:
##  $ X                               : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Gender                          : Factor w/ 2 levels "Female","Male":
2 2 1 1 2 1 2 1 1 2 ...
##  $ customer_type                   : Factor w/ 2 levels "disloyal
Customer",..: 2 1 2 2 2 2 2 2 2 1 ...
##  $ age                             : int  13 25 26 25 61 26 47 52 41 20
...
##  $ type_of_travel                  : Factor w/ 2 levels "Business
travel",..: 2 1 1 1 1 2 2 1 1 1 ...
##  $ customer_class                  : Factor w/ 3 levels
"Business","Eco",..: 3 1 1 1 1 2 2 1 1 2 ...
##  $ flight_distance                 : int  460 235 1142 562 214 1180 1276
2035 853 1061 ...
##  $ inflight_wifi_service           : Factor w/ 6 levels
"0","1","2","3",..: 4 4 3 3 4 4 3 5 2 4 ...
##  $ departure_arrival_time_convenient: int  4 2 2 5 3 4 4 3 2 3 ...
```

```
##  $ ease_of_online_booking           : int  3 3 2 5 3 2 2 4 2 3 ...
##  $ gate_location                    : int  1 3 2 5 3 1 3 4 2 4 ...
##  $ food_and_drink                   : int  5 1 5 2 4 1 2 5 4 2 ...
##  $ online_boarding                  : int  3 3 5 2 5 2 2 5 3 3 ...
##  $ seat_comfort                     : int  5 1 5 2 5 1 2 5 3 3 ...
##  $ inflight_entertainment           : Factor w/ 6 levels
"0","1","2","3",..: 6 2 6 3 4 2 3 6 2 3 ...
##  $ onboard_service                  : int  4 1 4 2 3 3 3 5 1 2 ...
##  $ leg_room_service                 : int  3 5 3 5 4 4 3 5 2 3 ...
##  $ baggage_handling                 : int  4 3 4 3 4 4 4 5 1 4 ...
##  $ checkin_service                  : int  4 1 4 1 3 4 3 4 4 4 ...
##  $ inflight_service                 : int  5 4 4 4 3 4 5 5 1 3 ...
##  $ cleanliness                      : int  5 1 5 2 3 1 2 4 2 2 ...
##  $ departure_delay_in_minutes       : int  25 1 0 11 0 0 9 4 0 0 ...
##  $ arrival_delay_in_minutes         : num  18 6 0 9 0 0 23 0 0 0 ...
##  $ satisfaction                     : Factor w/ 2 levels "neutral or
dissatisfied",..: 1 1 2 1 2 1 1 2 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:393] 214 1125 1530 2005 2109
2486 2631 3622 4042 4491 ...
##   ..- attr(*, "names")= chr [1:393] "214" "1125" "1530" "2005" ...

library(ggplot2)
ggplot(airlinedata,aes(age, colour =satisfaction))+geom_freqpoly(binwidth=1)
```




Plotted the count of passengers that were neutral or dissatisfied vs the ampunt of satisfied passengers by age. Younger people were more likely to be dissatisfied.

```
set.seed(123)
airline.train.index = createDataPartition(y=airlinedata[,1],p=0.5)[[1]]
airline.train = airlinedata[airline.train.index,]
airline.valid<-airlinedata[-airline.train.index,]
summary(airline.train)
```

```
##        X             Gender                customer_type        age
##  Min.   :     1   Female:32965   disloyal Customer:11980   Min.   : 7.00
##  1st Qu.: 32456   Male  :31779   Loyal Customer   :52764   1st Qu.:27.00
##  Median : 64939                                           Median :40.00
##  Mean   : 64962                                           Mean   :39.46
##  3rd Qu.: 97408                                           3rd Qu.:51.00
##  Max.   :129878                                           Max.   :85.00
##         type_of_travel   customer_class  flight_distance
inflight_wifi_service
##  Business travel:44740   Business:30999   Min.   :  31    0: 1968
##  Personal Travel:20004   Eco     :29170   1st Qu.: 413    1:11150
##                          Eco Plus: 4575   Median : 843    2:16097
##                                           Mean   :1190    3:16122
##                                           3rd Qu.:1739    4:12301
##                                           Max.   :4983    5: 7106
```

```
##  departure_arrival_time_convenient ease_of_online_booking gate_location
##  Min.   :0.000                     Min.   :0.000          Min.   :0.000
##  1st Qu.:2.000                     1st Qu.:2.000          1st Qu.:2.000
##  Median :3.000                     Median :3.000          Median :3.000
##  Mean   :3.063                     Mean   :2.753          Mean   :2.978
##  3rd Qu.:4.000                     3rd Qu.:4.000          3rd Qu.:4.000
##  Max.   :5.000                     Max.   :5.000          Max.   :5.000
##  food_and_drink  online_boarding  seat_comfort   inflight_entertainment
##  Min.   :0.000   Min.   :0.00    Min.   :1.000   0:    4
##  1st Qu.:2.000   1st Qu.:2.00    1st Qu.:2.000   1: 7809
##  Median :3.000   Median :3.00    Median :4.000   2:11038
##  Mean   :3.201   Mean   :3.25    Mean   :3.445   3:11956
##  3rd Qu.:4.000   3rd Qu.:4.00    3rd Qu.:5.000   4:18207
##  Max.   :5.000   Max.   :5.00    Max.   :5.000   5:15730
##  onboard_service leg_room_service baggage_handling checkin_service
##  Min.   :1.000   Min.   :0.000    Min.   :1.00     Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000    1st Qu.:3.00     1st Qu.:3.000
##  Median :4.000   Median :4.000    Median :4.00     Median :3.000
##  Mean   :3.387   Mean   :3.348    Mean   :3.63     Mean   :3.312
##  3rd Qu.:4.000   3rd Qu.:4.000    3rd Qu.:5.00     3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000    Max.   :5.00     Max.   :5.000
##  inflight_service  cleanliness    departure_delay_in_minutes
##  Min.   :1.000    Min.   :0.000   Min.   :   0.00
##  1st Qu.:3.000    1st Qu.:2.000   1st Qu.:   0.00
##  Median :4.000    Median :3.000   Median :   0.00
##  Mean   :3.645    Mean   :3.288   Mean   :  14.45
##  3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:  12.00
##  Max.   :5.000    Max.   :5.000   Max.   :1592.00
##  arrival_delay_in_minutes                  satisfaction
##  Min.   :   0.00          neutral or dissatisfied:36620
##  1st Qu.:   0.00          satisfied              :28124
##  Median :   0.00
##  Mean   :  14.93
##  3rd Qu.:  13.00
##  Max.   :1584.00

summary(airline.valid)

##        X             Gender                customer_type        age
##  Min.   :     0   Female:32738   disloyal Customer:11734   Min.   : 7.0
##  1st Qu.: 32455   Male  :32005   Loyal Customer   :53009   1st Qu.:27.0
##  Median : 64937                                           Median :40.0
##  Mean   : 64910                                           Mean   :39.4
##  3rd Qu.: 97408                                           3rd Qu.:51.0
##  Max.   :129879                                           Max.   :85.0
##         type_of_travel   customer_class  flight_distance
inflight_wifi_service
##  Business travel:44705   Business:30991   Min.   :  31    0: 1940
##  Personal Travel:20038   Eco     :28947   1st Qu.: 414    1:11100
##                          Eco Plus: 4805   Median : 844    2:16139
```

```
##                                          Mean   :1191    3:15965
##                                          3rd Qu.:1744    4:12401
##                                          Max.   :4983    5: 7198
##  departure_arrival_time_convenient ease_of_online_booking gate_location
##  Min.   :0.000                     Min.   :0.00           Min.   :1.000
##  1st Qu.:2.000                     1st Qu.:2.00           1st Qu.:2.000
##  Median :3.000                     Median :3.00           Median :3.000
##  Mean   :3.052                     Mean   :2.76           Mean   :2.976
##  3rd Qu.:4.000                     3rd Qu.:4.00           3rd Qu.:4.000
##  Max.   :5.000                     Max.   :5.00           Max.   :5.000
##  food_and_drink  online_boarding  seat_comfort   inflight_entertainment
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   0:   14
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1: 7825
##  Median :3.000   Median :3.000   Median :4.000   2:10859
##  Mean   :3.209   Mean   :3.255   Mean   :3.438   3:11849
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5.000   4:18475
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   5:15721
##  onboard_service leg_room_service baggage_handling checkin_service
##  Min.   :0.00    Min.   :0.000    Min.   :1.000    Min.   :0.0
##  1st Qu.:2.00    1st Qu.:2.000    1st Qu.:3.000    1st Qu.:2.0
##  Median :4.00    Median :4.000    Median :4.000    Median :3.0
##  Mean   :3.38    Mean   :3.354    Mean   :3.634    Mean   :3.3
##  3rd Qu.:4.00    3rd Qu.:4.000    3rd Qu.:5.000    3rd Qu.:4.0
##  Max.   :5.00    Max.   :5.000    Max.   :5.000    Max.   :5.0
##  inflight_service  cleanliness    departure_delay_in_minutes
##  Min.   :0.000    Min.   :0.000   Min.   :   0.00
##  1st Qu.:3.000    1st Qu.:2.000   1st Qu.:   0.00
##  Median :4.000    Median :3.000   Median :   0.00
##  Mean   :3.639    Mean   :3.284   Mean   :  14.84
##  3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:  12.00
##  Max.   :5.000    Max.   :5.000   Max.   :1305.00
##  arrival_delay_in_minutes                  satisfaction
##  Min.   :   0.00          neutral or dissatisfied:36605
##  1st Qu.:   0.00          satisfied              :28138
##  Median :   0.00
##  Mean   :  15.26
##  3rd Qu.:  13.00
##  Max.   :1280.00
```

Partitioned the data into 50% training data and 50% validation data.

```
ftable(airline.train$satisfaction,airline.train$inflight_entertainment)

##                              0     1     2     3     4     5
##
## neutral or dissatisfied      4  6690  8724  8662  7066  5474
## satisfied                    0  1119  2314  3294 11141 10256
```

Created a pivot table of the inflight entertainment survey results based on customer satisfaction.

```
ftable(airline.train$satisfaction,airline.train$inflight_wifi_service)

##                               0     1     2     3     4     5
##
## neutral or dissatisfied       4  7455 12152 12003  4928    78
## satisfied                  1964  3695  3945  4119  7373  7028
```

Created a pivot table of the in flight WiFi service survey results based on customer satisfaction.

People that did not use the WiFi service were primarily still satisfied overall. More passengers that rated the WiFi service poorly ended up being dissatisfied overall. There is a very small amount of people that rated the WiFi service a 5 and ended up being dissatisfied overall.

```
round(prop.table(table(airline.train$satisfaction,airline.train$inflight_ente
rtainment),margin=1),2)

##
##                               0    1    2    3    4    5
##   neutral or dissatisfied 0.00 0.18 0.24 0.24 0.19 0.15
##   satisfied               0.00 0.04 0.08 0.12 0.40 0.36

round(prop.table(table(airline.train$satisfaction,airline.train$inflight_wifi
_service),margin=1),2)

##
##                               0    1    2    3    4    5
##   neutral or dissatisfied 0.00 0.20 0.33 0.33 0.13 0.00
##   satisfied               0.07 0.13 0.14 0.15 0.26 0.25
```

Created a pivot table for each service that shows the probabilities.

A lower rating given for in flight WiFi service was more likely to result in overall dissatisfaction.

```
variables<-c(8,15,24)
set.seed(123)
airline.train.index2 =
createDataPartition(airlinedata$satisfaction,p=0.5,list=FALSE)
airline.train2 = airlinedata[airline.train.index2,variables]
airline.valid2<-airlinedata[-airline.train.index2,variables]
```

Partitioned the data again using only the 3 necessary variables.

```
round(prop.table(table(airline.train2$satisfaction,airline.train2$inflight_en
tertainment),margin=1),2)

##
##                               0    1    2    3    4    5
##   neutral or dissatisfied 0.00 0.18 0.23 0.24 0.19 0.15
##   satisfied               0.00 0.04 0.08 0.12 0.40 0.36
```

```
round(prop.table(table(airline.train2$satisfaction,airline.train2$inflight_wi
fi_service),margin=1),2)

##
##                              0    1    2    3    4    5
##   neutral or dissatisfied 0.00 0.20 0.33 0.33 0.13 0.00
##   satisfied               0.07 0.13 0.14 0.14 0.26 0.25
```

Created two pivot tables of this data partition to show the probabilities of each outcome.

```
airline.nb<-naiveBayes(airline.train2$satisfaction ~ .,data = airline.train2)
airline.nb

##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
## neutral or dissatisfied               satisfied
##               0.5655041               0.4344959
##
## Conditional probabilities:
##                          inflight_wifi_service
## Y                                         0           1           2
3
##   neutral or dissatisfied 0.0001911889 0.2047633354 0.3281894409
0.3307022096
##   satisfied               0.0699584089 0.1328427713 0.1421563400
0.1431161352
##                          inflight_wifi_service
## Y                                         4           5
##   neutral or dissatisfied 0.1343238740 0.0018299511
##   satisfied               0.2603533468 0.2515729978
##
##                          inflight_entertainment
## Y                                         0           1           2
3
##   neutral or dissatisfied 0.0003277524 0.1843880589 0.2334416737
0.2377297681
##   satisfied               0.0000000000 0.0397426327 0.0842131456
0.1152465252
##                          inflight_entertainment
## Y                                         4           5
##   neutral or dissatisfied 0.1940840685 0.1500286783
##   satisfied               0.3976396147 0.3631580818
```

Computed the Naive Bayes probability.

The probability that a customer will be neutral or dissatisfied is 57%. The probability that a customer will be satisfied is 43%.