Logistic Regression

Khang Thai, David Park, Jonathan Ho, David Favela

Reading in the Dataset

```
Invistico_Airline <- read.csv("Invistico_Airline.csv", header=TRUE)
str(Invistico_Airline)
```

```
## 'data.frame':    129880 obs. of  23 variables:
##  $ satisfaction                     : chr  "satisfied" "satisfied" "satisfied" "satisfied"
...
##  $ Gender                           : chr  "Female" "Male" "Female" "Female" ...
##  $ Customer.Type                    : chr  "Loyal Customer" "Loyal Customer" "Loyal Customer"
"Loyal Customer" ...
##  $ Age                              : int  65 47 15 60 70 30 66 10 56 22 ...
##  $ Type.of.Travel                   : chr  "Personal Travel" "Personal Travel" "Personal Trav
el" "Personal Travel" ...
##  $ Class                            : chr  "Eco" "Business" "Eco" "Eco" ...
##  $ Flight.Distance                  : int  265 2464 2138 623 354 1894 227 1812 73 1556 ...
##  $ Seat.comfort                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Departure.Arrival.time.convenient: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Food.and.drink                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Gate.location                    : int  2 3 3 3 3 3 3 3 3 3 ...
##  $ Inflight.wifi.service            : int  2 0 2 3 4 2 2 2 5 2 ...
##  $ Inflight.entertainment           : int  4 2 0 4 3 0 5 0 3 0 ...
##  $ Online.support                   : int  2 2 2 3 4 2 5 2 5 2 ...
##  $ Ease.of.Online.booking           : int  3 3 2 1 2 2 5 2 4 2 ...
##  $ On.board.service                 : int  3 4 3 1 2 5 5 3 4 2 ...
##  $ Leg.room.service                 : int  0 4 3 0 0 4 0 3 0 4 ...
##  $ Baggage.handling                 : int  3 4 4 1 2 5 5 4 1 5 ...
##  $ Checkin.service                  : int  5 2 4 4 4 5 5 5 5 3 ...
##  $ Cleanliness                      : int  3 3 4 1 2 4 5 4 4 4 ...
##  $ Online.boarding                  : int  2 2 2 3 5 2 3 2 4 2 ...
##  $ Departure.Delay.in.Minutes       : int  0 310 0 0 0 0 17 0 0 30 ...
##  $ Arrival.Delay.in.Minutes         : int  0 305 0 0 0 0 15 0 0 26 ...
```

Convert satisfaction into a factor

```
Invistico_Airline$satisfaction <- as.factor(Invistico_Airline$satisfaction)
```

Create New Columns Rating Sum & Rating Mean

```
Invistico_Airline$ratingSum <- as.numeric(apply(Invistico_Airline[,8:21], 1, sum))
Invistico_Airline$ratingMean <- c(Invistico_Airline$ratingSum/14)
```

Create train & test sets

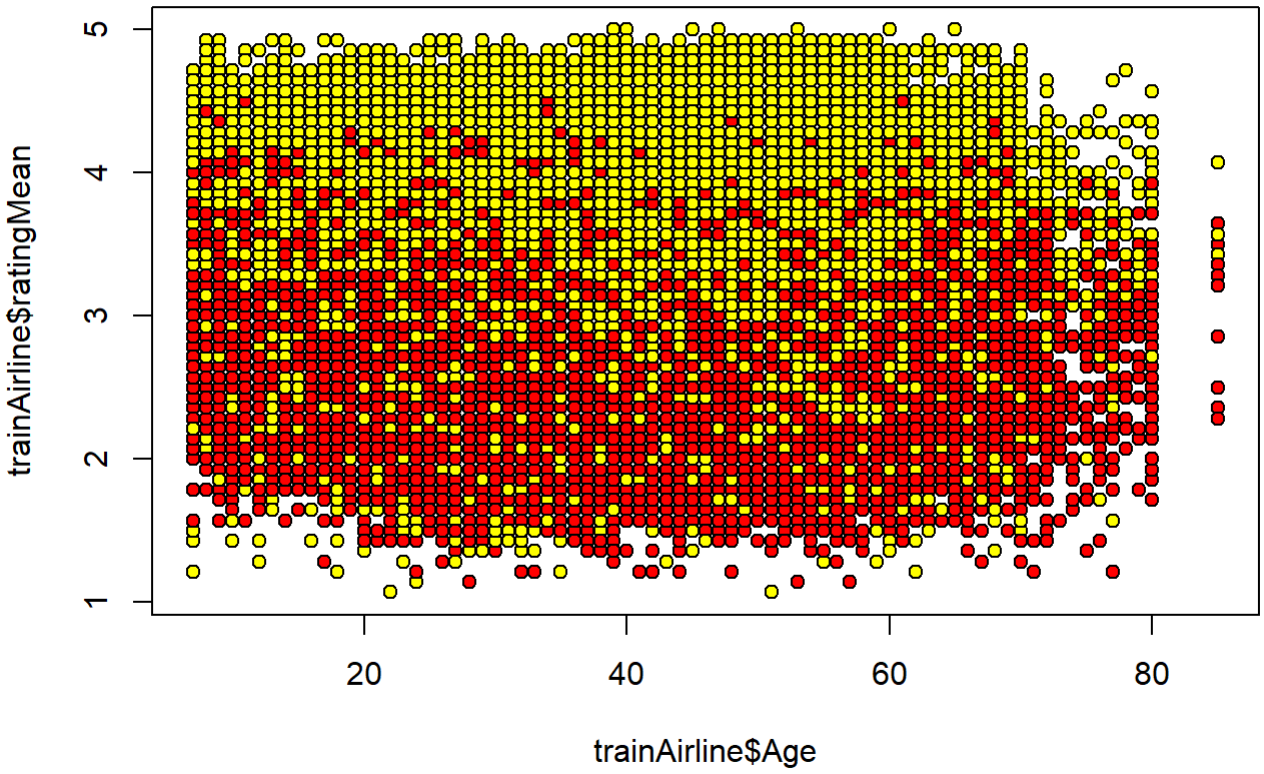
```
set.seed(3)
i <- sample(1:nrow(Invistico_Airline), 0.8*nrow(Invistico_Airline), replace = FALSE)
trainAirline <- Invistico_Airline[i,]
testAirline <- Invistico_Airline[-i,]
```

Clean out columns not needed

```
trainAirline <- trainAirline[,c(4,7,25,1)]
testAirline <- testAirline[,c(4,7,25,1)]
```

Plotting the dataset based on Age and RatingMean

```
plot(trainAirline$Age, trainAirline$ratingMean, pch=21, bg=c("red","yellow")
     [unclass(trainAirline$satisfaction)])
```



Create a Logistic Regression Model

```
glm1 <- glm(satisfaction~., data=trainAirline, family=binomial)
summary(glm1)
```

```
##
## Call:
## glm(formula = satisfaction ~ ., family = binomial, data = trainAirline)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5220  -0.9097   0.4168   0.8560   2.9525
##
## Coefficients:
##                    Estimate Std. Error  z value Pr(>|z|)
## (Intercept)     -6.652e+00  5.268e-02 -126.268  < 2e-16 ***
## Age              1.290e-02  4.945e-04   26.080  < 2e-16 ***
## Flight.Distance -3.032e-05  7.417e-06   -4.088 4.35e-05 ***
## ratingMean       1.948e+00  1.363e-02  142.931  < 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: 143083  on 103903  degrees of freedom
## Residual deviance: 112235  on 103900  degrees of freedom
## AIC: 112243
##
## Number of Fisher Scoring iterations: 4
```

Test the Logistic Model

```
probs <- predict(glm1, newdata = testAirline, type="response")
pred <- ifelse(probs>0.5,2,1)
acc1 <- mean(pred==as.integer(testAirline$satisfaction))
print(paste("glm1 accuracy = ", acc1))
```

```
## [1] "glm1 accuracy =  0.731752386818602"
```

```
table(pred, as.integer(testAirline$satisfaction))
```

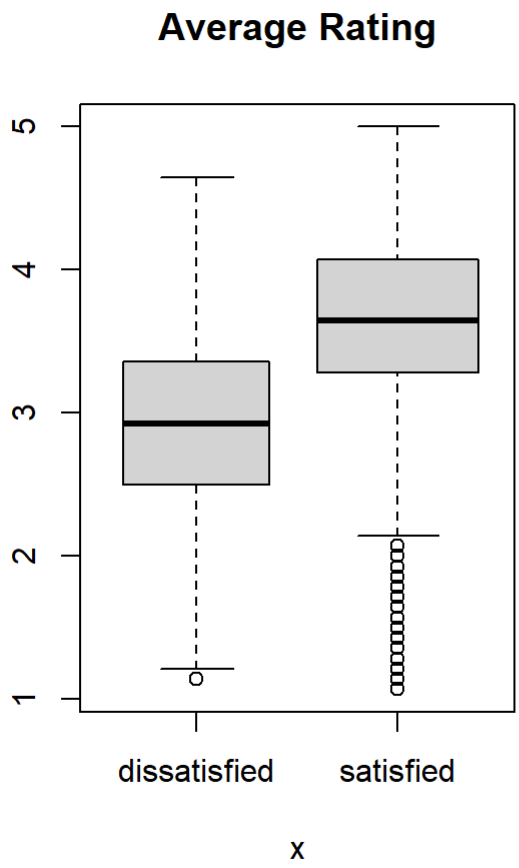
```
##
## pred     1     2
##    1  7916  3057
##    2  3911 11092
```

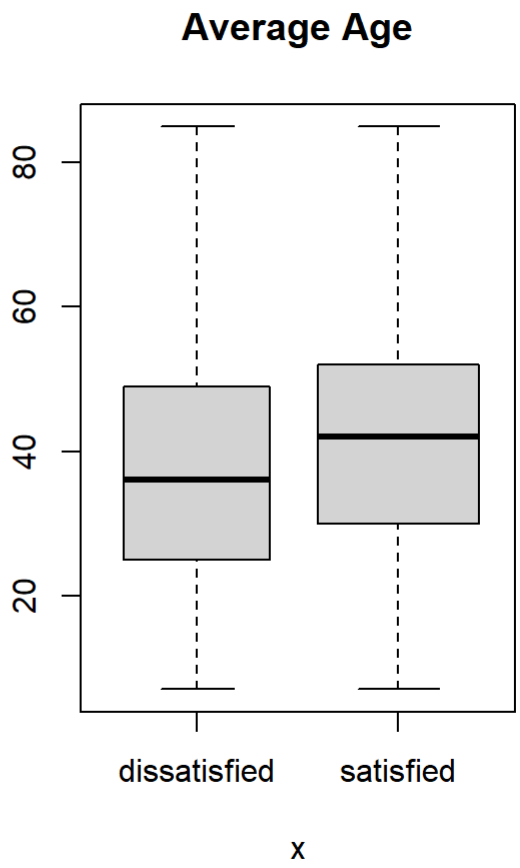
row 1 = total satisfied row 2 = total unsatisfied

row 1, col 1 = True Pos row 1, col 2 = False Pos row 2, col 1 = False Neg row 2, col 2 = True Neg

Graph for average rating depending on if it the user was satisfied or not

```
par(mfrow=c(1,2))
plot(trainAirline$satisfaction, trainAirline$ratingMean, main="Average Rating", ylab="", varwidt
h=TRUE)
plot(trainAirline$satisfaction, trainAirline$Age, main="Average Age", ylab="", varwidth=TRUE)
```





knn Classification

```
library(class)
str(trainAirline)
```

```
## 'data.frame':    103904 obs. of  4 variables:
##  $ Age            : int  34 39 41 44 39 49 26 39 41 28 ...
##  $ Flight.Distance: int  2481 2600 1594 1767 2023 951 1970 261 2267 2323 ...
##  $ ratingMean     : num  2.86 4.14 3.29 3.14 2.64 ...
##  $ satisfaction   : Factor w/ 2 levels "dissatisfied",..: 1 2 2 2 1 2 1 2 2 1 ...
```

```
airline_pred <- knn(train=trainAirline[,c(1,3)], test = testAirline[,c(1,3)], cl = trainAirline
[,4], k=3)
```

Predicting Satisfaction and finding the True/False Positives and Negatives

```
results <- airline_pred == testAirline$satisfaction
acc <- length(which(results==TRUE))/ length(results)
table(results, airline_pred)
```

```
##        airline_pred
## results dissatisfied satisfied
##   FALSE         3393      3531
##   TRUE          8296     10756
```

```
acc
```

```
## [1] 0.7334463
```

Decision Tree Regression

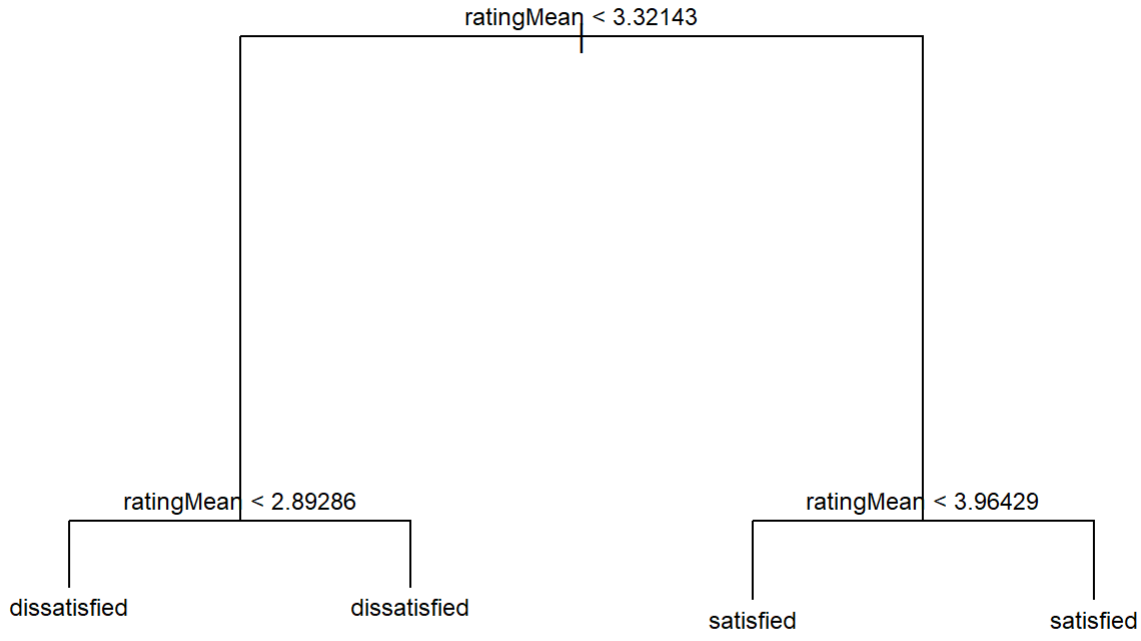
```
library(tree)
tree_airline <- tree(satisfaction~., data=trainAirline)
tree_airline
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
## 1) root 103904 143100 satisfied ( 0.45201 0.54799 )
##   2) ratingMean < 3.32143 51077  63930 dissatisfied ( 0.68134 0.31866 )
##     4) ratingMean < 2.89286 28334  29700 dissatisfied ( 0.78213 0.21787 ) *
##     5) ratingMean > 2.89286 22743  31240 dissatisfied ( 0.55578 0.44422 ) *
##   3) ratingMean > 3.32143 52827  57010 satisfied ( 0.23028 0.76972 )
##     6) ratingMean < 3.96429 34201  42150 satisfied ( 0.30645 0.69355 ) *
##     7) ratingMean > 3.96429 18626  11310 satisfied ( 0.09041 0.90959 ) *
```

```
summary(tree_airline)
```

```
##
## Classification tree:
## tree(formula = satisfaction ~ ., data = trainAirline)
## Variables actually used in tree construction:
## [1] "ratingMean"
## Number of terminal nodes:  4
## Residual mean deviance:  1.101 = 114400 / 103900
## Misclassification error rate: 0.2737 = 28441 / 103904
```

```
plot(tree_airline)
text(tree_airline, cex=0.75, pretty = 0)
```




```
pred <- predict(tree_airline, newdata = testAirline, type = "class")
table(pred, testAirline$satisfaction)
```

```
##
## pred           dissatisfied satisfied
##   dissatisfied         8810      4043
##   satisfied            3017     10106
```

```
mean(pred==testAirline$satisfaction)
```

```
## [1] 0.7282107
```

Comparing the results

The logistic regression model gave us a 0.7311 accuracy on the test data used. The KNN model created gave us slightly higher 0.7334 accuracy. The models were created using the airline data and the same train and test data sets were used. Looking into the tables, we can see that the KNN model has slightly lower true satisfied

predictions but had higher true dissatisfied predictions, making it slightly more accurate than the logistic regression model.

How Results Were Achieved?

The logistic regression model created weights and adjusts them as data is being fed and the algorithm is run. From the statistic graphs that we made, we can clearly see there are some differences in the data that will contribute to the model, such as more dissatisfied uses giving lower ratings and those that are of older age being more satisfied with their airline. The KNN model decides depending on the proximity and relation with the neighboring points, and the values those points have. The graph demonstrating the plots of satisfied and unsatisfied uses that has rating as a Y-label and age as an X-label displays a trend that most people that are dissatisfied give out lower ratings. This hypothesis was also proven as we got a higher than 0.7 accuracy.