TEXAS TECH UNIVERSITY
Rawls College *of* Business

DATA AND TEXT MINING

US Airline Passenger Satisfaction

KRISHNA CHARAN BAJANTHRI CHIKALAGURIKI

CONTENTS

INTRODUCTION

In the aviation industry, high-grade customer satisfaction is a key factor to run the business, as the airline industry is very competitive and customer loyalty varies with small changes in the services. Therefore, companies need to understand the customers' need to deliver unparalleled experiences to retain customers.

Using the customer's satisfaction data obtained from Kaggle, we here attempt to understand the reasons for customer experience being satisfied or not. Based on that, improvements will be made to provide better service by the airline company. Also, as part of the analysis, we will be able to understand several factors which improve customer satisfaction level.

In this era of social networking, the customers get to Twitter, Facebook, Blogs or Google Reviews to review or share their experience on the internet. Any miserable experience of a customer, which is shared becomes viral on the internet, seriously affecting the brand or goodwill of the company.

To arrest these unwanted viral trends on the internet and to avoid heavy legal compensations or penalties, the companies in the service industry are identifying a customer's satisfaction through a service rating cards or survey after delivering service and compensating a dissatisfied customer upon the fault of company service. This rewards or compensation program has effectively reduced the complaints raised by a customer.

PROBLEM STATEMENT

In our scenario, the airline company wants to identify a customer satisfaction level, based on his rating on various aspects of airline experience. Hence, primarily we build a model to classify the customer satisfaction level. Precisely we will classify customers' being satisfied or not and accordingly try to find out the factors related to high satisfaction.

DATA ABSTRACT

The data we obtained from Kaggle represents US airline satisfaction data consisting of 129880 observations and 23 attributes. The Satisfaction level, which is our dependent variable, is represented as a factor ("Satisfied" and "Neutral or Dissatisfied").

Variable	Variable Description	Variable Value Level
Satisfaction (Dependent Var)	Airline satisfaction level	Satisfaction, neutral or dissatisfaction
Age	The actual age of the passengers	
Gender	Gender of the passengers	Female, Male
Type of Travel	Purpose of the flight of the passengers	Personal Travel, Business Travel
Class	Travel class in the plane of the passengers	Business, Eco, Eco Plus
Customer Type	The customer type	Loyal customer, disloyal customer
Flight distance	The flight distance of this journey	

Inflight wifi service	Satisfaction level of the inflight wifi service	Rating: 0 (least) - 5 (highest)
Ease of Online booking	Satisfaction level of online booking	Rating: 0 (least) - 5 (highest)
Inflight service	Satisfaction level of inflight service	Rating: 0 (least) - 5 (highest)
Online boarding	Satisfaction level of online boarding	Rating: 0 (least) - 5 (highest)
Inflight entertainment	Satisfaction level of inflight entertainment	Rating: 0 (least) - 5 (highest)
Food and drink	Satisfaction level of Food and drink	Rating: 0 (least) - 5 (highest)
Seat comfort	Satisfaction level of Seat comfort	Rating: 0 (least) - 5 (highest)
On-board service	Satisfaction level of On-board service	Rating: 0 (least) - 5 (highest)
Leg room service	Satisfaction level of Leg room service	Rating: 0 (least) - 5 (highest)
Departure/Arrival time convenient	Satisfaction level of Departure/Arrival time convenient	Rating: 0 (least) - 5 (highest)
Baggage handling	Satisfaction level of baggage handling	Rating: 0 (least) - 5 (highest)
Gate location	Satisfaction level of Gate location	Rating: 0 (least) - 5 (highest)
Cleanliness	Satisfaction level of Cleanliness	Rating: 0 (least) - 5 (highest)
Check-in service	Satisfaction level of Check-in service	Rating: 0 (least) - 5 (highest)
Departure Delay in Minutes	Minutes delayed when departure	
Arrival Delay in Minutes	Minutes delayed when Arrival	
Flight cancelled	Whether the Flight cancelled or not	Yes, No
Flight time in minutes	Minutes of Flight takes	

Data Source: https://www.kaggle.com/johnddddd/customer-satisfaction

DATA PREPARATION

We import the dataset into the R environment and observe the structure of the data loaded.

```
## 'data.frame':    129880 obs. of  24 variables:
##  $ ï..id                           : int  11112 110278 103199 47462 12001
1 100744 32838 32864 53786 7243 ...
##  $ satisfaction_v2                 : Factor w/ 2 levels "neutral or dissa
tisfied",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Gender                          : Factor w/ 2 levels "Female","Male":
1 2 1 1 1 2 1 2 1 2 ...
##  $ Customer.Type                   : Factor w/ 2 levels "disloyal Custome
r",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Age                             : int  65 47 15 60 70 30 66 10 56 22
 ...
##  $ Type.of.Travel                  : Factor w/ 2 levels "Business travel
",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Class                           : Factor w/ 3 levels "Business","Eco
",..: 2 1 2 2 2 2 2 2 1 2 ...
##  $ 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 ...
```

We observe the dataset having 129880 observations and 24 attributes. The attribute "satisfaction_v2" represents the satisfaction level of the customer on two different levels: "neutral or dissatisfied" and "satisfied". Also, we drop the ID attribute as we wouldn't require that in our analysis or model building. Hence, we have 1 dependent variable and 22

independent variables in our dataset. We check our dataset for any possible NA values, which must be dealt before we proceed with building the model.

```
##          Arrival.Delay.in.Minutes
##                               393
```

We observe 363 NA values in the attribute “Arrival.Delay.in.Minutes”. Let us analyze the dataset further to understand whether the NA values are truly random or have any implied reason behind. To perform this, we initially create a new dataset, which contains the NA values. Then we compare the mean values of the attributes of this dataset with the original dataset mean attribute values to observe any significant difference.Following we observe the mean values of attributes of original dataset and the dataset with NA values.

```
##                                    mean_data mean_na_data
## satisfaction_v2                     1.547328     1.521628
## Gender                              1.492616     1.501272
## Customer.Type                       1.816908     1.832061
## Age                                39.427957    39.162850
## Type.of.Travel                      1.309416     1.368957
## Class                               1.593864     1.646310
## Flight.Distance                  1981.409055  2113.229008
## Seat.comfort                        2.838597     2.842239
## Departure.Arrival.time.convenient   2.990645     3.111959
## Food.and.drink                      2.851994     2.842239
## Gate.location                       2.990422     3.005089
## Inflight.wifi.service               3.249130     3.239186
## Inflight.entertainment              3.383477     3.295165
## Online.support                      3.519703     3.432570
## Ease.of.Online.booking              3.472105     3.450382
## On.board.service                    3.465075     3.442748
## Leg.room.service                    3.485902     3.414758
## Baggage.handling                    3.695673     3.765903
## Checkin.service                     3.340807     3.366412
## Cleanliness                         3.705759     3.664122
## Online.boarding                     3.352587     3.366412
## Departure.Delay.in.Minutes         14.713713    37.885496
## Arrival.Delay.in.Minutes           15.091129           NA
```

By comparing the mean values of different attributes, we observe that the attributes have same values and imply that the NA values are truly random. Hence, we can replace the NA values with mean values of the attribute or omit the missing, however before proceeding any further, let us analyze the “Arrival.Delay.in.Minutes” variable’s correlation with other variables.

```
##                                    Arrival.Delay.in.Minutes
## satisfaction_v2                                      -0.081
## Gender                                                0.001
## Customer.Type                                        -0.005
## Age                                                  -0.011
## Type.of.Travel                                       -0.006
## Class                                                 0.014
## Flight.Distance                                       0.110
## Seat.comfort                                         -0.026
## Departure.Arrival.time.convenient                     0.003
## Food.and.drink                                       -0.015
## Gate.location                                         0.004
## Inflight.wifi.service                                -0.035
## Inflight.entertainment                               -0.033
## Online.support                                       -0.036
## Ease.of.Online.booking                               -0.040
## On.board.service                                     -0.041
## Leg.room.service                                      0.000
## Baggage.handling                                     -0.014
## Checkin.service                                      -0.024
## Cleanliness                                          -0.067
## Online.boarding                                      -0.022
## Departure.Delay.in.Minutes                            0.965
```

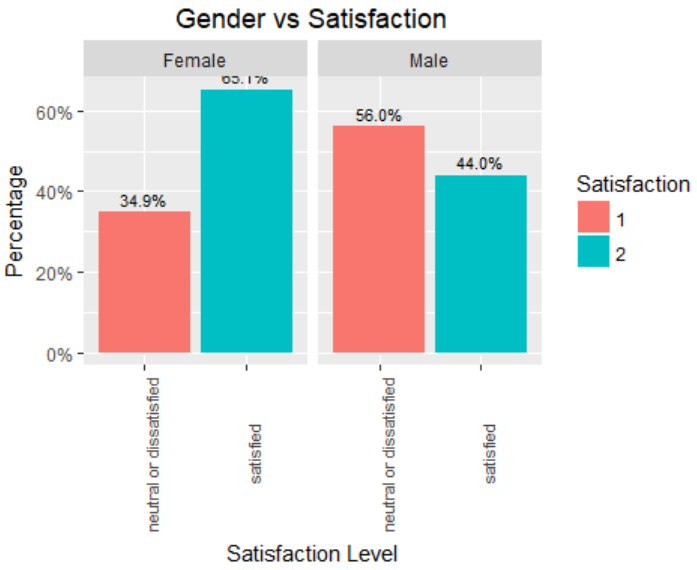
We observe a very high correlation (0.965) between Departure.Delay.in.Minutes and Arrival.Delay.in.Minutes. Fortunately, without the need of replacing or omitting the NA values in the variable “Arrival.Delay.in.Minutes”, we can exclude the variable entirely, as we have a similar variable having no NA values in our dataset to proceed with our modeling.

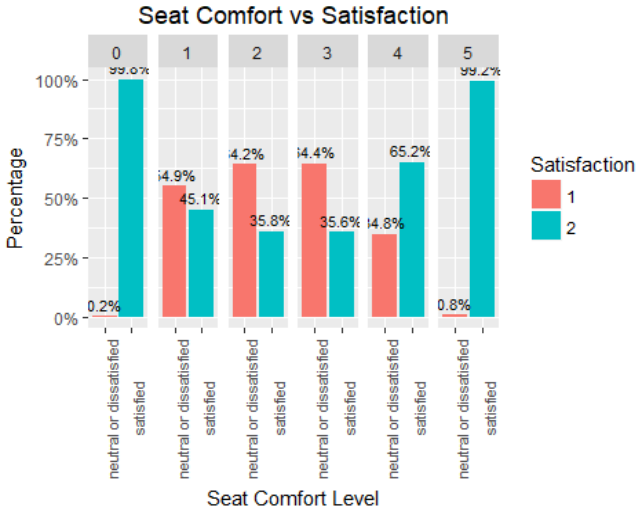
EXPLORATORY ANALYSIS

After observing the impact of several independent variables over the “Satisfaction level”, we present below few visualizations displaying a level of satisfaction at different factor levels for the variables “Gender”, “Class”, “Seat Comfort”, “Inflight Entertainment”, “Online Support” and “Baggage Handling”.

On an inferential perspective, these variables are considered the most significant in terms of airline customer satisfaction, which is more likely the case as we observe the following visualizations.

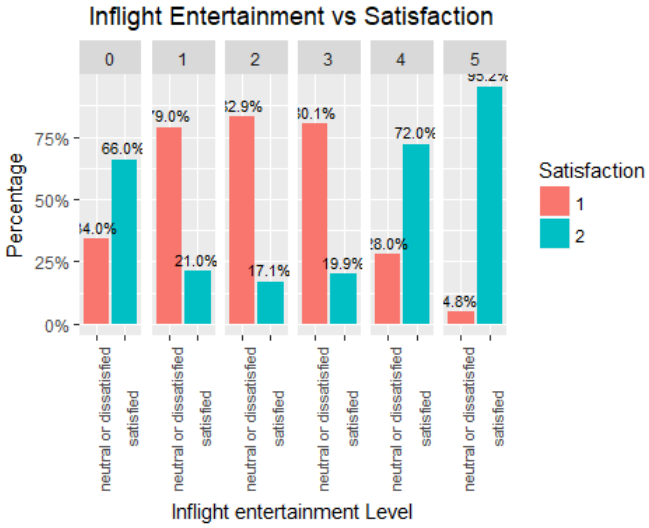
We observe that the Female customers are comparatively more satisfied than the Male customers as we can see 65% female customers are satisfied against 34% dissatisfied.





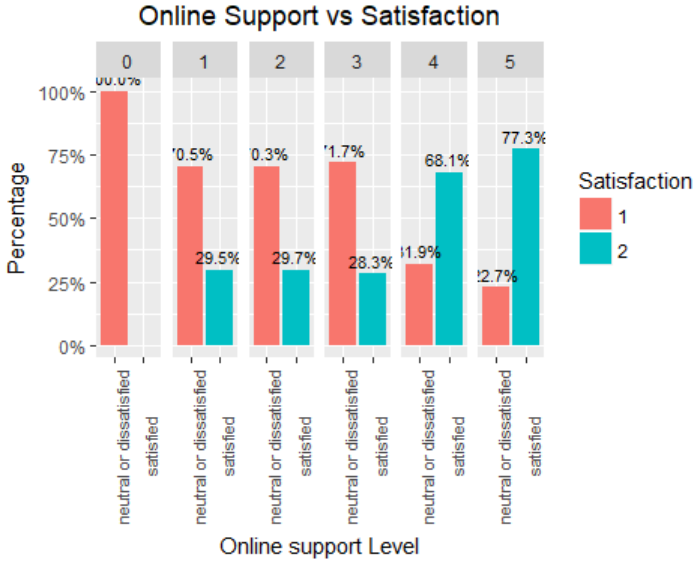
We observe that Seat Comfort is having a significant effect on the customer satisfaction level, as we see that customers rating 5 on seat comfort are 99 percent satisfied.

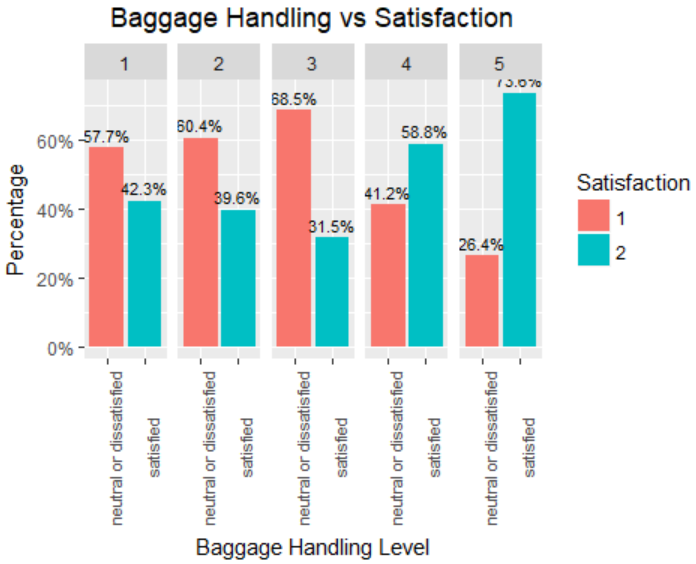
We observe that the customers traveling the Business Class are more satisfied than the customers traveling in Economy classes.



We observe that Inflight Entertainment is having a significant effect on the customer satisfaction level, as we see that customers rating 5 on seat comfort are 96 percent satisfied.

We observe that Online Support is having a significant effect on the customer satisfaction level, as we see that customers rating 0 on seat comfort are 99 percent dissatisfied.





We observe that Baggage Handling is having a significant effect on the customer satisfaction level, as we see that customers rating 5 on seat comfort are 74 percent satisfied.

DATA MODELLING

We build models to predict the satisfaction level of the customer based on the independent variables of our dataset. This will help the airline company to understand the satisfaction level of its customers and hence take necessary actions, in order to improve customer satisfaction. Here we build 3 models to predict the satisfaction level, namely Logistic regression model, Linear discriminant analysis and Decision tree models.

DATA PROCESSING AND NORMALIZATION

Initially, we check for the class bias to ensure our models are better.

```
##
## neutral or dissatisfied                satisfied
##                    45.3                     54.7
```

We observe there is no class bias, as both the cases are approximately equally distributed in the dataset.

Before we begin modeling the dataset, we convert the factorial data in the dataset into numeric type as we require in the model building and further analysis, except for our dependent variable.

As an initial step in the dataset, we observe few of the variables in our dataset are on the different scale and hence mislead our modeling techniques. Hence, we normalize the data.

TRAINING AND TEST DATASETS

Now that we have the normalized dataset, we divide the dataset into training and testing dataset in the ratio of 70:30. This will help us in understanding the performance of our model.

MODELS

1. LOGISTIC REGRESSION

To build a classification model to predict satisfaction, we start with the go-to classification model: Logistic Regression model. Here, we utilize the Logistic regression classification algorithm from the GLM package in R to predict "satisfaction_v2" from the set of independent variables available.

```
## glm(formula = satisfaction_v2 ~ ., family = "binomial", data = data_train)
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -6.89178    0.07364 -93.588  < 2e-16 ***
## Gender                           -0.96808    0.01958 -49.447  < 2e-16 ***
## Customer.Type                     2.01801    0.02934  68.777  < 2e-16 ***
## Age                              -0.59347    0.05303 -11.191  < 2e-16 ***
## Type.of.Travel                   -0.89373    0.02602 -34.349  < 2e-16 ***
## Class                            -0.99828    0.03633 -27.482  < 2e-16 ***
## Flight.Distance                  -0.78212    0.07056 -11.084  < 2e-16 ***
## Seat.comfort                      1.36640    0.05476  24.952  < 2e-16 ***
## Departure.Arrival.time.convenient -0.96689    0.04027 -24.011  < 2e-16 ***
## Food.and.drink                   -1.05742    0.05573 -18.974  < 2e-16 ***
## Gate.location                     0.55810    0.04579  12.187  < 2e-16 ***
```

```
## Inflight.wifi.service             -0.37433    0.05283  -7.085 1.39e-12 ***
## Inflight.entertainment             3.48705    0.04940  70.591  < 2e-16 ***
## Online.support                     0.44155    0.05346   8.259  < 2e-16 ***
## Ease.of.Online.booking             1.09970    0.06902  15.932  < 2e-16 ***
## On.board.service                   1.53559    0.04902  31.324  < 2e-16 ***
## Leg.room.service                   1.09322    0.04173  26.196  < 2e-16 ***
## Baggage.handling                   0.42010    0.04445   9.452  < 2e-16 ***
## Checkin.service                    1.55768    0.04131  37.704  < 2e-16 ***
## Cleanliness                        0.43946    0.05737   7.661 1.85e-14 ***
## Online.boarding                    0.87812    0.05902  14.879  < 2e-16 ***
## Departure.Delay.in.Minutes        -7.89244    0.42188 -18.708  < 2e-16 ***

## Null deviance: 125222  on 90916  degrees of freedom
## Residual deviance:  70468  on 90895  degrees of freedom
## AIC: 70512
## Number of Fisher Scoring iterations: 5
```

From the above summary statistics of the model we built, we observe the coefficient weights and the importance of the variable from p-value. We observe that all the independent variables are significantly important in predicting the “satisfaction_v2”.

We shall predict the test dataset using the model built. Note that we obtained the predictions in terms of probabilities. Hence, by considering a cutoff of 0.5 (practically used default value), we classify the predicted “satisfaction_v2” into “Yes” or “No”. Let us build the confusion matrix, using the predicted and observed values to assess the performance of the model.

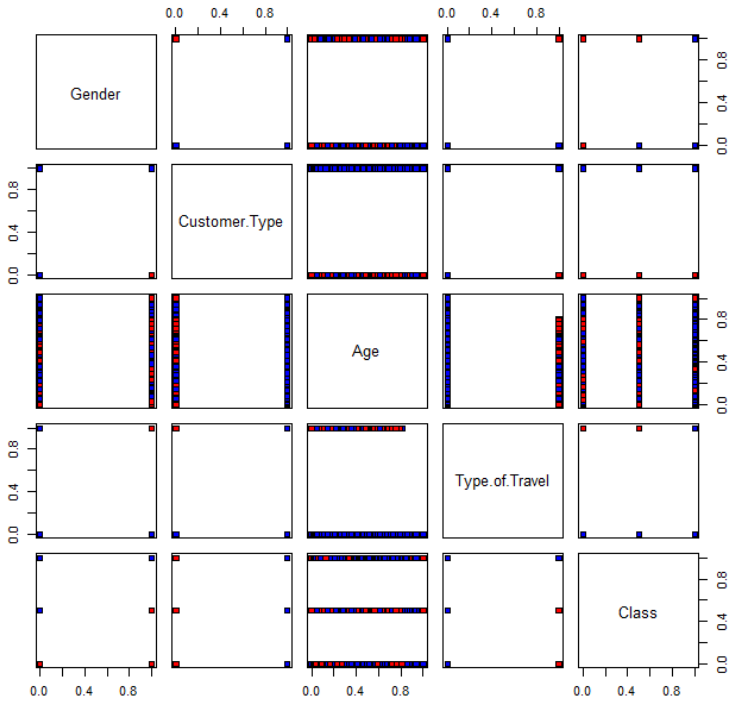
```
## Confusion Matrix and Statistics
##                          Reference
## Prediction                neutral or dissatisfied satisfied
##   neutral or dissatisfied                   14431      3171
##   satisfied                                  3206     18155
##                Accuracy : 0.8363
##                  95% CI : (0.8326, 0.84)
##     No Information Rate : 0.5473
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6696
##             Sensitivity : 0.8513
##             Specificity : 0.8182
##        'Positive' Class : satisfied
```

From the above results, we can observe that:

- The overall accuracy of 83.6% and Kappa value of 69%. That means, our model performed better than a random prediction of the dependent variable.
- Sensitivity value of 85 percent, indicating that our model has correctly identified 85% of the satisfied customers correctly.
- Specificity value of 81 percent, indicating that our model has correctly identified 81 percent of the “neutral or dissatisfied” customers correctly.

2. LINEAR DISCRIMINANT ANALYSIS

One significant difference we observe between Logistic Regression and LDA is that when the groups in the dataset are clearly classified logistic regression is unstable and hence LDA model provides a better model. Let us observe the predicted class separability.



As we can observe that the “satisfied” and “neutral or dissatisfied” classes are more likely separable, we do not expect to see much significance using the LDA model over the Logistic model. However, we build a model using the LDA.

We predict the test dataset using the LDA model we built and compute a confusion matrix to understand the performance of the model.

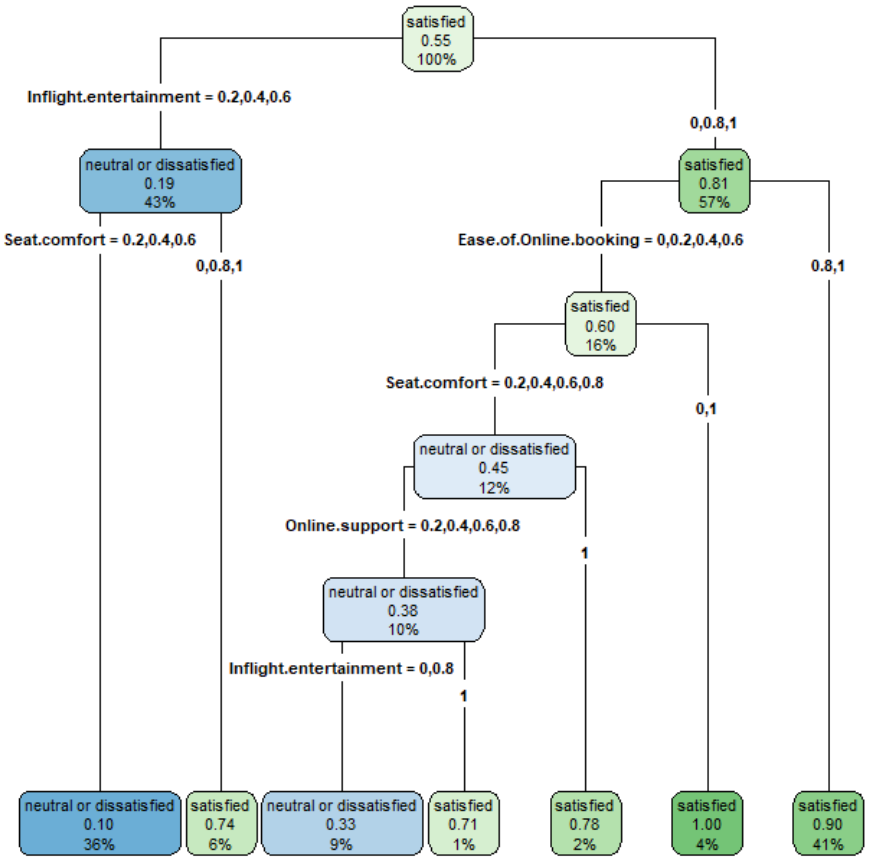
```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction                neutral or dissatisfied satisfied
##   neutral or dissatisfied                   14473      3164
##   satisfied                                  3151     18175
##
##                Accuracy : 0.8379
##                  95% CI : (0.8342, 0.8416)
##     No Information Rate : 0.5477
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6729


##             Sensitivity : 0.8517
##             Specificity : 0.8212
##
##        'Positive' Class : satisfied
##
```

From the above results, we can observe that:

- The overall accuracy of 83.7% and Kappa value of 67%.
- Sensitivity value of 85 percent, indicating that our model has correctly identified 85% of the satisfied customers correctly.
- Specificity value of 82 percent, indicating that our model has correctly identified 81 percent of the “neutral or dissatisfied” customers correctly.

3. DECISION TREE – CART ALGORITHM

We use a decision tree algorithm, CART (Classification and Regression Tree) for classifying the dependent variable based on the set of independent variables. This algorithm repeatedly partitions the data into multiple-subsets so that the leaf nodes are pure or homogeneous. Here we ensure that our categorical variables are in factor form so that the CART algorithm uses classification tree rather than regression tree. Following we visualize the model we built.



The algorithm built a classification tree, as we observe that the first or the more valuable variable we have “Inflight entertainment” being rated 1-3 we classify them to be “neutral or dissatisfied” class on left node, which is further classified based on the rating given on “seat

comfort". If seat comfort is rated 1-3, we finally decided them to be "neutral or dissatisfied", else classified them to be "satisfied". Similarly, the tree progresses finally into 7 leaf nodes. Note that the factor levels are normalized (Normalized Value (Rating Value): 0(0), 0.2(1), 0.4(2), 0.6(3), 0.8(4), 1(5)).

We use this model to predict the test data set and observe the confusion matrix to understand the model performance.

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction                neutral or dissatisfied satisfied
##   neutral or dissatisfied                   15053      2478
##   satisfied                                  2584     18848
##
##                Accuracy : 0.8701
##                  95% CI : (0.8667, 0.8734)
##     No Information Rate : 0.5473
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.7377
##             Sensitivity : 0.8848
##             Specificity : 0.8535
##
##        'Positive' Class : satisfied
```

From the above results, we can observe that:

- The overall accuracy of 87% and Kappa value of 73%. That means, our model performed better than a random prediction of the dependent variable.
- Sensitivity value of 88 percent, indicating that our model has correctly identified 88% of the satisfied customers correctly.
- Specificity value of 85 percent, indicating that our model has correctly identified 85 percent of the "neutral or dissatisfied" customers correctly.

MODEL COMPARISON

In order to compare the model performance of the 3 models we built on a common metrics, we extract the accuracy, Kappa statistic, Sensitivity and Specificity from the confusionMatrix() output.

	Logistic Regression	Linear Discriminant Analysis	Decision Tree
Accuracy	0.8363	0.8379	0.8700
Kappa	0.6696	0.6728	0.7376
Sensitivity	0.8513	0.8517	0.8838
Specificity	0.8182	0.8212	0.8535

We observe from the above summary that,

- Logistic regression and LDA models we built provide similar performance metrics for our dataset.
- Decision tree provides better performance metrics over the other two models we built.

Below we compare the models one against the other.

1. LDA OVER LOGISTIC REGRESSION MODEL

As we expected, we do not see significant improvement in the model performance, as our predicted class was clearly more likely separable, and we have sufficient predictors. However, we only see a slight 1 percent improvement in the Specificity of our model, which represents correctly identifying “neutral or dissatisfied” customers. For our business model, it is critical to identify the negative class correctly, as the airline company is looking forward to identify the unsatisfied customers and reward them to improve their satisfaction.

2. DECISION TREE OVER LDA

We obtained a decent 5 percent improvement in the model accuracy using the decision tree algorithm over the logistic regression and LDA models we built. Also, we obtained a 5 percent improvement in the specificity of our model which is a critical metric for our business model.

CONCLUSION

We conclude that decision tree model we built using the CART algorithm, is statistically better suitable for our dataset and business model providing us with an accuracy of 87 percent along with 88 percent correctly identifying “satisfied” customers (Sensitivity) and 85 percent correctly identifying “neutral or dissatisfied” customers (Specificity).

Also, we observed from exploratory analysis and decision tree as well, the Inflight experience and Seat Comfort level and Online service level significantly affect the customer experience along with several other variables considered. The airline service companies must ensure high quality of service in these parameters to ensure a high level of customer satisfaction.

The models we built did lack to perfectly classify customer satisfaction level. The prediction accuracy can be further improved using advanced classification algorithms or deep learning concepts.

REFERENCES

1. Lecture Notes – Data and Text Mining
2. Lantz, B. (2015). Machine learning with R: Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R. Birmingham: PACKT Publishing.

APPENDIX – R SOURCE CODE

```
#Loading Libraries

library(dplyr)
library(caret)
library(rpart)
library(rpart.plot)
library(e1071)
library(MASS)
library(corrplot)
library(ggplot2)
library(tidyr)

#Loading data

data <- read.csv("C:/Education/Data and Text Mining/Project/Data/satisfaction.csv")
str(data)

#drop ID attribute
data <- data[-1]

#Dealing with NA values

#NA Count
sapply(data, function(x) sum(is.na(x)))

#NA dataset
data.na <- data %>% filter(is.na(data$Arrival.Delay.in.Minutes))

#Mean values for attributes in Original Dataset
a<-as.data.frame(c())
d<-colnames(data)
b<-c()
for(i in 1:ncol(data)){
  if(class(data[,i])=="factor"){
    b[i] <- mean(as.numeric(data[,i]))
  }
  else{
    b[i] <- mean(data[,i], na.rm = T)

  }
}
names(b) <- colnames(data)
a <- as.data.frame(b)

#Mean values for attributes in NA dataset
d<-c()
```

```
e <- as.data.frame(c())
for(i in 1:ncol(data.na)){
  if(class(data.na[,i])=="factor"){
    d[i] <- mean(as.numeric(data.na[,i]))
  }
  else{
    d[i] <- mean(data.na[,i])
      }
}
names(d) <- colnames(data.na)
e <- as.data.frame(d)

f <- cbind(a, e)
colnames(f) <- c("mean_data", "mean_na_data")
f

data.non <- data %>% filter(!is.na(data$Arrival.Delay.in.Minutes))
for(i in 1:ncol(data.non)){
  if(class(data[,i])!="integer"){
    data.non[,i] <- as.integer(data.non[,i])
  }
}
data.cor <- round(cor(data.non[,1:22], data.non[,23]),3)
colnames(data.cor) <- "Arrival.Delay.in.Minutes"
data.cor

data <- data[-23]
```

```
#Exploratory Analysis
```

```
ggplot(data, aes(satisfaction_v2, group = Gender)) +
          geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count
") +
          geom_text(aes( label = scales::percent(..prop..),y= ..prop..
 ), stat= "count", vjust = -.5, cex=3) +
          labs(x="Satisfaction Level",y = "Percentage",title="Gender vs
 Satisfaction", fill="Satisfaction") +
          scale_y_continuous(labels=scales::percent) +
          facet_grid(~Gender)+ theme(axis.text.x = element_text(size=8,
angle = 90), plot.title = element_text(hjust = 0.5))

ggplot(data, aes(satisfaction_v2, group = Class)) +
          geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count
") +
          geom_text(aes( label = scales::percent(..prop..),y= ..prop..
 ), stat= "count", vjust = -.5, cex=3) +
          labs(x="Class of Travel",y = "Percentage",title="Travel Class
 vs Satisfaction", fill="Satisfaction") +
          scale_y_continuous(labels=scales::percent) +
```

```
          facet_grid(~Class)+ theme(axis.text.x = element_text(size=8,a
ngle = 90), plot.title = element_text(hjust = 0.5))

ggplot(data, aes(satisfaction_v2, group = Seat.comfort)) +
          geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count
") +
          geom_text(aes( label = scales::percent(..prop..),y= ..prop..
 ), stat= "count", vjust = -.5, cex=3) +
          labs(x="Seat Comfort Level",y = "Percentage",title="Seat Comf
ort vs Satisfaction", fill="Satisfaction") +
          scale_y_continuous(labels=scales::percent) +
          facet_grid(~Seat.comfort)+ theme(axis.text.x = element_text(s
ize=8,angle = 90), plot.title = element_text(hjust = 0.5))

ggplot(data, aes(satisfaction_v2, group = Inflight.entertainment)) +
          geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count
") +
          geom_text(aes( label = scales::percent(..prop..),y= ..prop..
 ), stat= "count", vjust = -.5, cex=3) +
          labs(x="Inflight entertainment Level",y = "Percentage",title=
"Inflight Entertainment vs Satisfaction", fill="Satisfaction") +
          scale_y_continuous(labels=scales::percent) +
          facet_grid(~Inflight.entertainment)+ theme(axis.text.x = elem
ent_text(size=8,angle = 90), plot.title = element_text(hjust = 0.5))

ggplot(data, aes(satisfaction_v2, group = Online.support)) +
          geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count
") +
          geom_text(aes( label = scales::percent(..prop..),y= ..prop..
 ), stat= "count", vjust = -.5, cex=3) +
          labs(x="Online support Level",y = "Percentage",title="Online
Support vs Satisfaction", fill="Satisfaction") +
          scale_y_continuous(labels=scales::percent) +
          facet_grid(~Online.support)+ theme(axis.text.x = element_text
(size=8,angle = 90), plot.title = element_text(hjust = 0.5))

ggplot(data, aes(satisfaction_v2, group = Baggage.handling)) +
          geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count
") +
          geom_text(aes( label = scales::percent(..prop..),y= ..prop..
 ), stat= "count", vjust = -.5, cex=3) +
          labs(x="Baggage Handling Level",y = "Percentage",title="Bagga
ge Handling vs Satisfaction", fill="Satisfaction") +
          scale_y_continuous(labels=scales::percent) +
          facet_grid(~Baggage.handling)+ theme(axis.text.x = element_te
xt(size=8,angle = 90), plot.title = element_text(hjust = 0.5))
```

```
# Data Modelling
```

```
round(prop.table(table(data$satisfaction_v2))*100,digit=1)

for(i in 2:ncol(data)){
  if(class(data[,i])=="factor"){
    data[,i] <- as.integer(data[,i])
  }
}
normalize <- function(x){
  return ((x-min(x))/(max(x)-min(x)))
}
data_n <- as.data.frame(lapply(data[2:22], normalize))
data_n <- cbind(data_n, satisfaction_v2=data$satisfaction_v2)

set.seed(5)
partition <- createDataPartition(y = data_n$satisfaction_v2, p=0.7, lis
t = F)
data_train <- data_n[partition,]
data_test <- data_n[-partition,]
```

```
# Logistic regression
```

```
log_model <- glm(satisfaction_v2~., data=data_train, family = "binomial
")
summary(log_model)

log_preds <- predict(log_model, data_test[,1:22], type = "response")
head(log_preds)

log_class <- array(c(99))
for (i in 1:length(log_preds)){
  if(log_preds[i]>0.5){
    log_class[i]<-"satisfied"
  }else{
    log_class[i]<-"neutral or dissatisfied"
  }
}
#Creating a new dataframe containing the actual and predicted values.
log_result <- data.frame(Actual = data_test$satisfaction_v2, Prediction
 = log_class)

mr1 <- confusionMatrix(as.factor(log_class), data_test$satisfaction_v2,
 positive = "satisfied")
mr1
```

```
# Linear Discriminant Analysis
```

```
pairs(data_train[,1:5], main="Predict ", pch=22,
   bg=c("red", "blue")[unclass(data_train$satisfaction_v2)])
```

```
lda_model <- lda(satisfaction_v2 ~ ., data_train)
lda_model

#Predict the model
lda_preds <- predict(lda_model, data_test)

mr2 <- confusionMatrix(data_test$satisfaction_v2, lda_preds$class, posi
tive = "satisfied")
mr2
```

```
#Decision tree - CART algorithm
```

```
data_train_new <- data_train
for(i in c(1,2,4,5,7,8,9,10,11,12,13,14,15,16,17,18,19,20)){
  data_train_new[,i] <- as.factor(data_train_new[,i])
}
data_test_new <- data_test
for(i in c(1,2,4,5,7,8,9,10,11,12,13,14,15,16,17,18,19,20)){
  data_test_new[,i] <- as.factor(data_test_new[,i])
}
cart_model <- rpart(satisfaction_v2 ~ ., data_train_new, method="class
")
cart_model
rpart.plot(cart_model,digits = 2,  type=4, extra=106)

cm1 <- predict(cart_model, data_test_new, type = "class")
mr3 <- confusionMatrix(cm1, data_test$satisfaction_v2, positive = "sati
sfied")
mr3
```

```
# Model Comparision
```

```
comp_cfm <- function(x1,x2,x3,n1,n2,n3){
    a <- data.frame()
    b <- as.matrix(c(x1[[3]][1:2],x1[[4]][1:2]))
    c <- as.matrix(c(x2[[3]][1:2],x2[[4]][1:2]))
    d <- as.matrix(c(x3[[3]][1:2],x3[[4]][1:2]))
    colnames(b) <- n1
    colnames(c) <- n2
    colnames(d) <- n3
    a <- as.data.frame(cbind(b,c,d))
    return(a)
}

r <- comp_cfm(mr1,mr2,mr3, "LogisticRegression", "LinearDiscriminantAna
lysis", "DecisionTree")
r
```