

Understanding Attributes That Improve Performances for Southeast

Airlines and their Regional Partners

Kushal Shah, Dishank Solanki, Yicun Deng, Roderick Cushing

Table of Contents

1. Introduction
2. Exploratory Data Analysis
 - a. Dataset Structure
 - b. Additions to the Dataset
 - c. Visualizations
3. Modelling
 - a. Association Rules
 - b. Support Vector Machines
 - c. Linear Regressions
4. Findings and Recommendations
 - a. Findings and Recommendations for FlyFast Airways
 - b. Findings and Recommendations for All Airlines
5. Conclusion

Introduction

In this paper we attempt to analyze survey data from Southeast Airlines customers. We look at each subsidiary airline partner to understand which airlines are performing well and which airlines are not performing well. We hope to provide insight with our analysis that can help improve the service of airlines that are not performing as well as others. Moreover, as there is always room for improvement, we also want to provide insight into why some airlines are performing well and use those insights to further boost their performance. By using a variety of machine learning techniques and by exploring the vast amount of data provided, we have created various recommendations that we believe will strongly strengthen the performance of Southeast Airlines and its regional partners.

Exploratory Data Analysis

Dataset Structure

The dataset is initially a JSON file that we convert to a functioning data frame that can be effectively used for an analysis in R. We achieve this by using the jsonlite package in R. Each row in the dataset represents one customer and the flight they took. The dataset consists of a variety of categorical and numerical variables that provide information about the customer, the airline, and the specific flight in question. We can classify the dataset columns into Customer Attributes, Flight and Airport Attributes and Travel Attributes. Certain variables can be considered a part of two groups. For example Amount Spent on Food and Dining could be considered a customer attribute and an airport attribute, as it is both the airport having popular restaurants and a customer willing to spend money. For simplicity, we assigned these types of variables to one group.

Customer Attributes Include:

1. Age
2. Gender
3. Year of first Flight
4. Type of Travel
5. Loyalty Status
6. Type of travel

Flight and Airport Attribute Include:

1. Destination
2. Origin
3. Money Spent on Shopping

4. Date of Travel
5. Airline
6. Scheduled Departure Hour
7. Money Spent on Food and Dining

Travel Attribute Include:

1. Flight Distance
2. Delay in Arrival
3. Delay in Departure
4. Flight time

Additions to the Dataset

We had a variety of different variables to the initial dataset to further improve our analysis and to further understand the data we were working with. To begin with, we incorporated different dummy variables. Dummy variables are essentially binary variables that will have the value of 1 if the condition is met and 0 if the condition is not met. In order to do so we used the fastDummies package available in R along with ifelse statements. A snippet of this code can be seen here:

```
data <- fastDummies::dummy_cols(survey, select_columns = "Gender") ##Creating Dummy Variables for the column Gender
data$Flight.cancelled <- ifelse(data$Flight.cancelled=="No",0,1) ##Creating a dummy variable for flights being
cancelled or not
```

We added dummy variables for the following variables:

1. Type of Travel (Business, Personal, Mileage)
2. Gender (Male, Female)
3. Promoters and Detractors
4. Class (Business, Economy, Economy+)

5. Long Trip

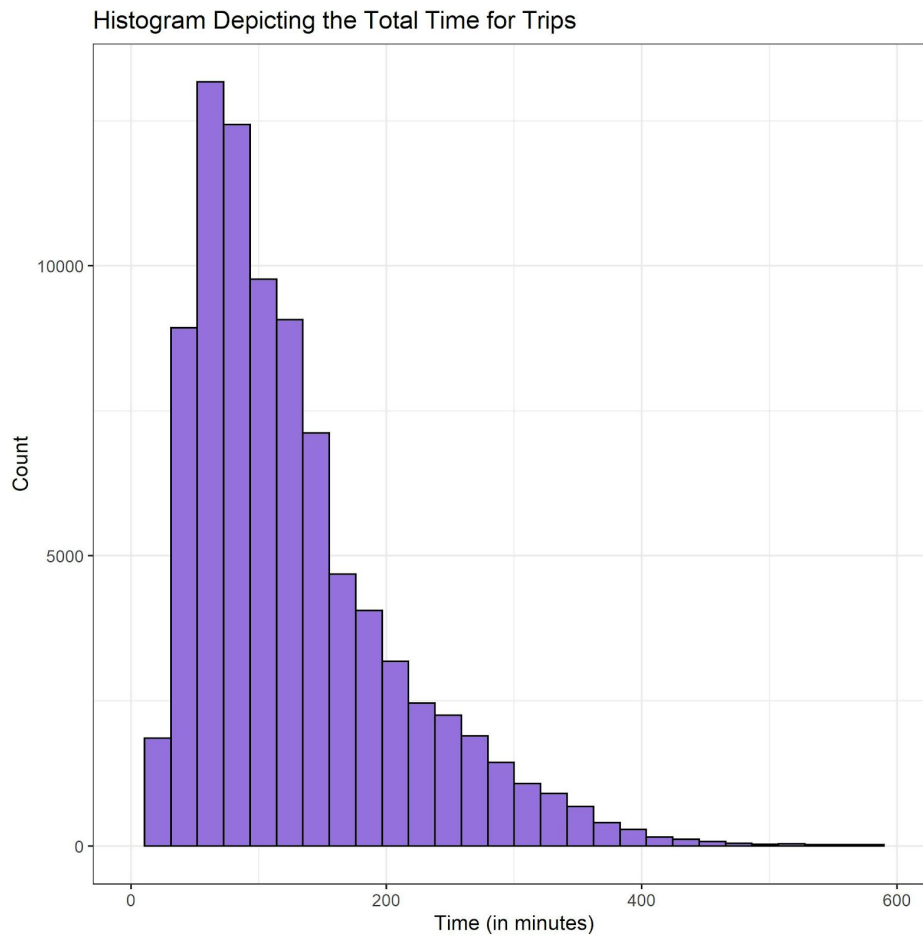
6. Flight Cancellation

In addition to the dummy variables we added other variables that are based on the existing variables. These variables are:

1. Total Trip Time = The Total Time of a Trip Including Delays.
2. Total Airport Expenditure = Total Amount Spent at the Airport

Here is a histogram (and the code for it) of this newly created variable to provide clarity on what the variable looks like in terms of its distribution:

```
TotalTripHist <- ggplot(data, aes(TotalTripTime)) + geom_histogram(color="black", fill="mediumpurple") + xlim(0,600) +  
ggtitle("Histogram Depicting the Total Time for Trips") + ylab("Count") + xlab("Time (in minutes)") + theme_bw()  
TotalTripHist
```



We can see the majority of trips are near 60 minutes (1 hour) long. An overwhelming majority of trips are less than 200 minutes long. The distribution is heavily skewed towards the left, implying among the range of times, most are closer to minimum than the maximum. This is further highlighted by the peak of the distribution being on the left side.

We also created various scales for certain variables to categorize similar values together. This was primarily used to categorize numerical data. The variable we categorized include:

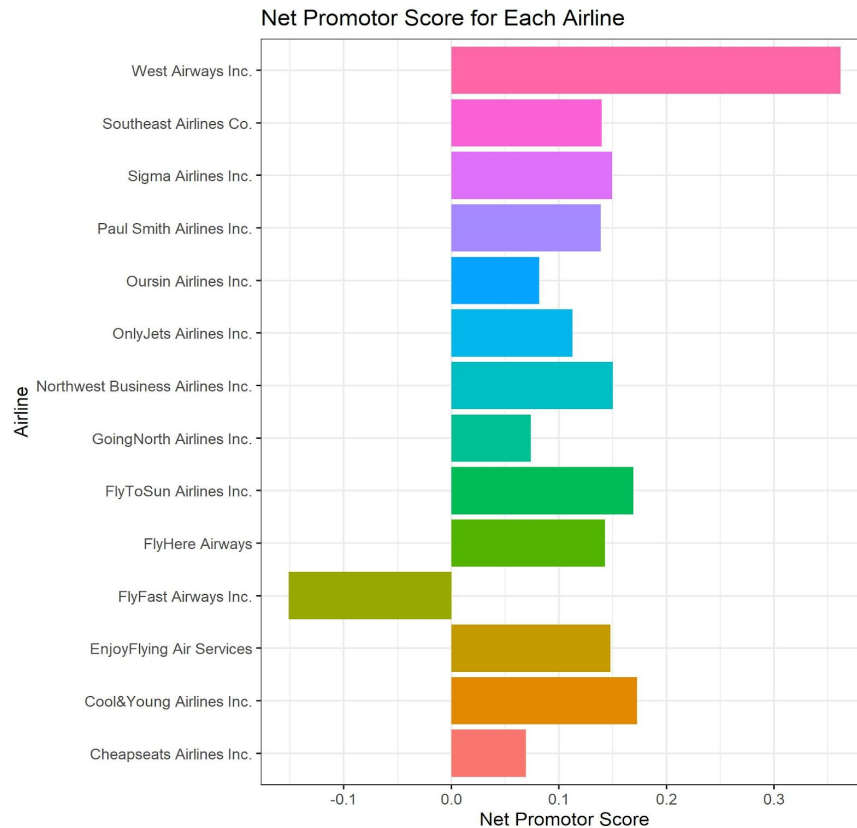
1. Flight Distance (Short, Medium, and Long)
2. Month (Start of the month, Middle of the month, and End of the month)
3. Scheduled Departure (Morning, Afternoon, Evening, and Night)
4. Likelihood to Recommend (Low, Medium, and High)
5. Age (Young, Adult, and Old)

Visualizations

In this section we continued our initial data analysis by creating different visualization. We explored a variety of the variables in a visual manner to further understand the dataset and determine which aspects we would like to explore. This included using boxplots, maps, and bar charts.

One of the first metrics we looked at was Net Promoter Score. We looked at each airline and their net promoter score to understand which airlines have a high net promoter score and which have a low net promoter score.

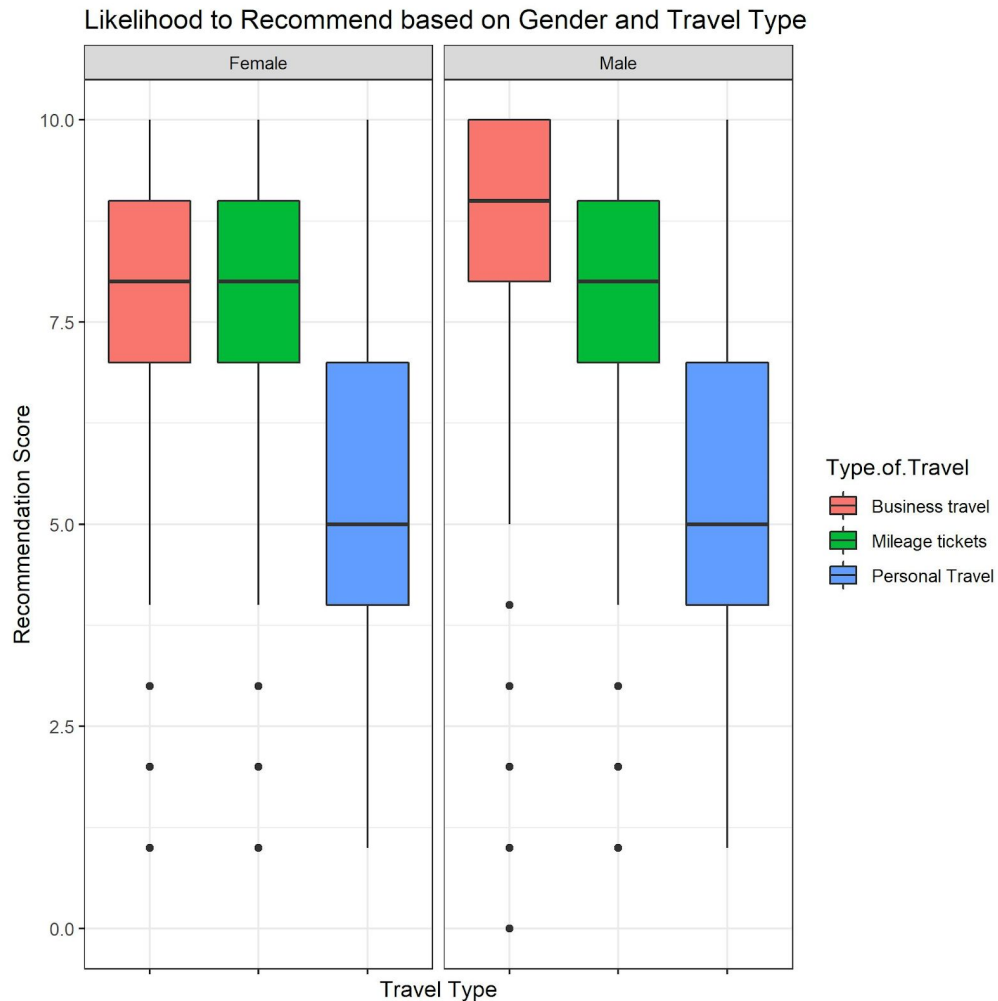
```
airlineNPS <- ggplot(airline, aes(x=Airline, y=NPS, fill=Airline)) + geom_bar(stat="identity") + coord_flip() +  
theme_bw() + theme(legend.position = "none") + ggtitle("Net Promotor Score for Each Airline") + xlab("Airline") +  
ylab("Net Promotor Score")  
airlineNPS
```



We can see that FlyFast airways clearly does not perform well, they are the only airline with a NPS that is negative. While a majority of the other airlines have a similar NPS, West Airways is clearly the best performing airline with a NPS significantly larger than every other airline.

We also looked into the relationship between the reason to travel with the recommendation score a customer would provide. This relationship is expressed using a box plot below, we also split the plots by gender. Hence with this graph we hope to isolate the type of customer and the reason for travel's impact on a recommendation score. The code used to create this visualization is also below:

```
TravelTypeScore <- ggplot(data, aes(x=Type.of.Travel, y=Likelihood.to.recommend, fill=Type.of.Travel)) + geom_boxplot()
+ theme_bw() + facet_wrap(~Gender) + theme(axis.text.x = element_blank()) + ylab("Recommendation Score") + xlab("Travel
Type") + ggtitle("Likelihood to Recommend based on Gender and Travel Type")
TravelTypeScore
```

It is clear from the graph that customers, both male and female, provide higher recommendation scores when travelling for business reasons. Travelling for personal reasons tend to lead to lower recommendation scores as well which is an important piece of information that can be used when deciding which audience to focus marketing on.

This next visualization is a map visualization that we would like to completely explain. We used the packages of ggmap and ggplot2 to create the visualization of. From origin city to destination city. The first step is to assemble all the departure and arrival city into a unique list:

```
mappingdata$Destination.City<-as.character(mappingdata$Destination.City)#convert des city into char
mappingdata$Origin.City<-as.character(mappingdata$Origin.City)#convert ori city into char
airports <- unique(c(mappingdata$Origin.City,mappingdata$Destination.City))#get the flight city
```

This is done in order to gain their geocode from google's api(they are requiring keys to access to their api now):

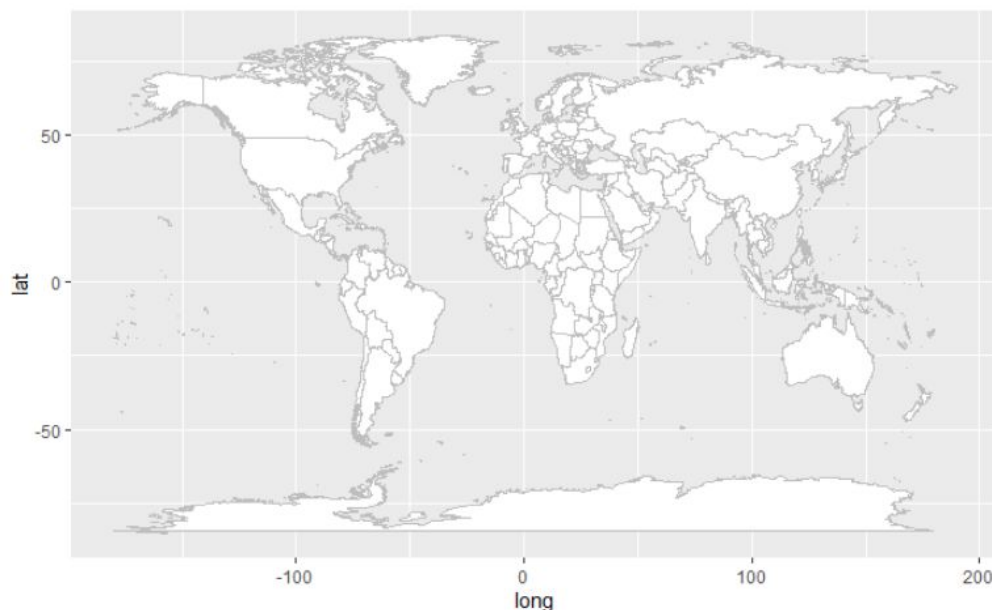
```
register_google(key = "AIzaSyBCVcJDlma8PiUI9qm878X7KZC4aoeZkhA", write = TRUE)#set the google api key
coords <- geocode(airports)#get the geocode of flight city
airports <- data.frame(airport=airports, coords)#create airport information
```

This is then merged without initial dataset using the Destination City, Origin City, and Airport:

```
mappingdata <- merge(mappingdata, airports, by.x="Destination.City", by.y="airport")#merge from
mappingdata <- merge(mappingdata, airports, by.x="Origin.City", by.y="airport")#merge to
```

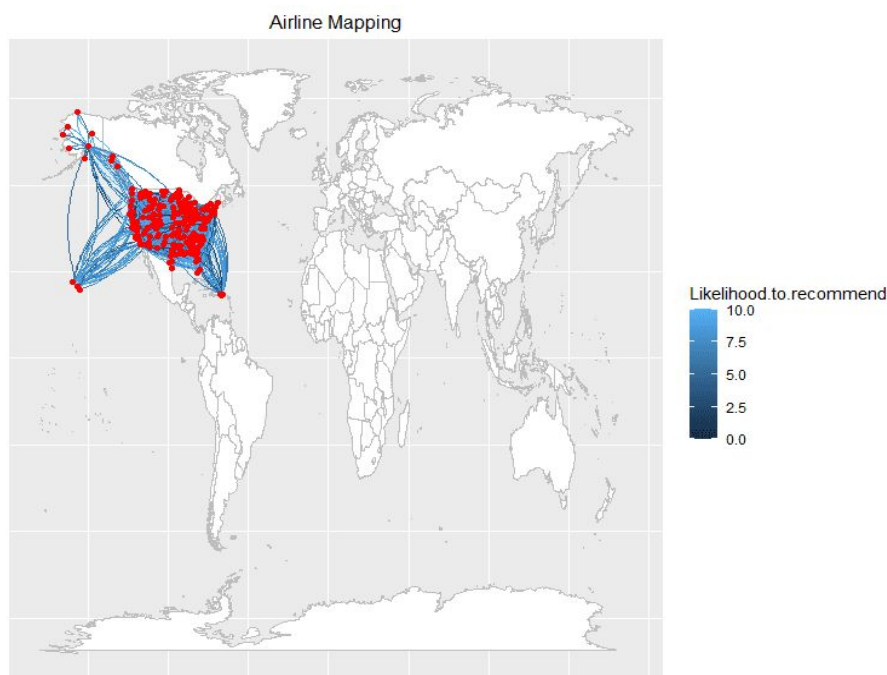
Then we create a layer of the World Map to plot the flight details. We do not use the US map as it looks confusing due to geographic locations of Hawaii, Alaska, and Puerto Rico. The code and what the map looks like can be seen below:

```
worldMap <- borders("world", colour="grey", fill="white")# create a layer of borders
```



Lastly, we create our plot. Within the plot we draw a curved line from the origin location to the destination location. This signifies the flight's travel. These lines are colored by the passenger's recommendation score. A darker line implies a higher score for that flight. The origin location of a flight is signified by a red dot. The code to create the final plot and the plot itself can be found below:

```
allUSA <- ggplot() + worldMap +  
  geom_curve(data=mappingdata,  
    aes(x=lon.y, y=lat.y, xend=lon.x, yend=lat.x, col=Likelihood.to.recommend),  
    size=.5,  
    curvature=0.2) +  
  geom_point(data=mappingdata,  
    aes(x=lon.y, y=lat.y),  
    colour="blue",  
    size=1.5) +  
  geom_point(data=mappingdata,  
    aes(x=lon.x, y=lat.x),  
    colour="red") +  
  theme(axis.line=element_blank(),  
    axis.text.x=element_blank(),  
    axis.text.y=element_blank(),  
    axis.title.x=element_blank(),  
    axis.title.y=element_blank(),  
    axis.ticks=element_blank(),  
    plot.title=element_text(hjust=0.5, size=12)) +  
  ggtitle("Airline Mapping")  
allUSA
```



We then focus on the United States and the final edited version of the visualization looks like this:



As we mentioned the lighter blue lines imply a lower recommendation score from customers and darker blue lines imply a higher recommendation score from customers. Flights between Florida and North East locations tend to have a lighter shade along with flights between Hawaii and the West Coast.

Modelling and Machine Learning

Based on our initial exploration of the data we identified two major aspects we would like to focus our modelling and analysis on:

1. Understanding Issues with FlyFast Airways and Identifying Areas for Improvement
2. Understanding Consumer and Flight Factors that Impact Recommendation Scores

While we strongly believe that FlyFast Airways required a large amount of attention, we wanted to make sure we also provided insight for the other airlines and what they can do to improve their performance. This is why we selected these 2 aspects to focus on.

Association Rules

We have used the Association rules to answer 2 business questions related to FlyFast Airways:

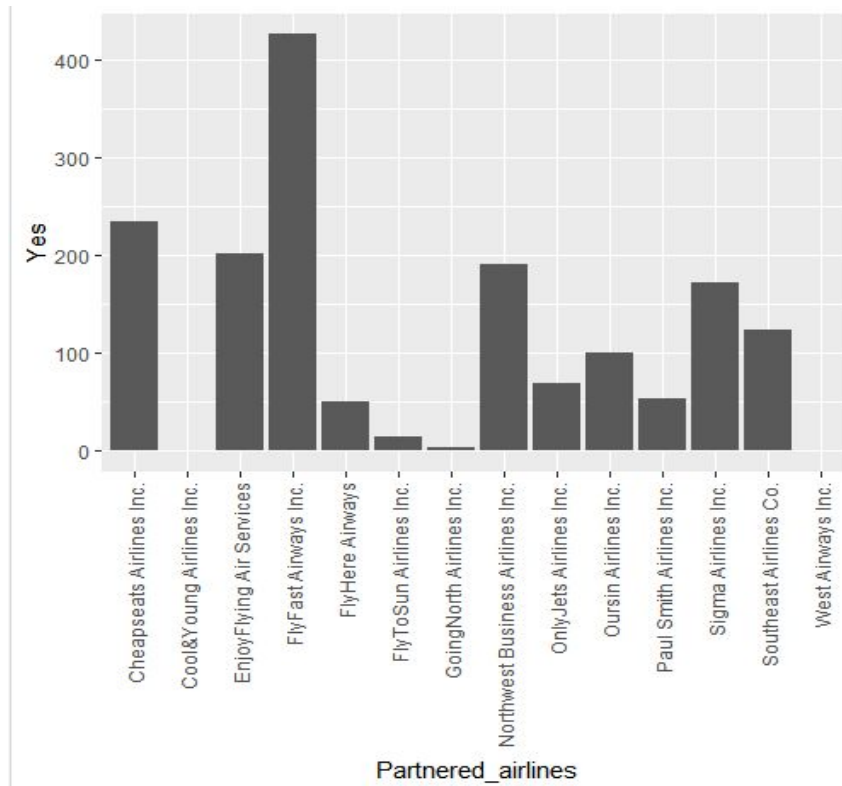
1. Rules to find cancellation of Flyfast Airways:

As FlyFast Airways had the lowest NPS score we decided to begin our analysis with Flyfast airways. We decided to find rules for cancellation as Flyfast Airways as they had the greatest number of cancelled flights. So we were curious to find which were the rules that defined why flights were getting cancelled for Flyfast Airways. We came to this conclusion

```
#####  
# visualizing partnered airlines for which most flights are cancelled  
#####  
  
table(survey$Partner.Name, survey$Flight.cancelled)  
survey_cancelled <- as.data.frame.matrix(table(survey$Partner.Name, survey$Flight.cancelled))  
view(survey_cancelled)  
survey_cancelled$Partnered_airlines <- rownames(survey_cancelled)  
ggplot(survey_cancelled, aes(y=Yes, x=Partnered_airlines)) +  
  geom_bar(stat="identity") + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

Through Bar graph visualization. Following is the snippet for bar graph:

The plot of bar graph is as follows:



Subset of data and data manipulation

Created the subset of the main Survey data frame named `survey_FlyFast_Yes` which has the data for partnered airlines FlyFast Airways and Flight.cancelled as Yes. As we have to find the rules for FlyFast Airways with Flight.Cancelled as Yes.

We identified 7 independent variables which were:

1. Airline.Status
2. Day.of.Month
3. Origin.State
4. Destination.State
5. Scheduled.Departure.Hour
6. Flight.Distance
7. Likelihood.to.recommend.

Our dependent variable was obviously Flight.Cancelled. The rest of the variables within the data frame were removed from the subset. As Apriori needs the variables into categories, converting the numeric column into categories:

Creating Flight.Distance.1 into category:

```
l
max(survey_FlyFast_Yes$Flight.Distance)
min(survey_FlyFast_Yes$Flight.Distance)
survey_FlyFast_Yes$Flight.Distance.1[survey_FlyFast_Yes$Flight.Distance>=73 & survey_FlyFast_Yes$Flight.Distance<=460] = 'Short'
survey_FlyFast_Yes$Flight.Distance.1[survey_FlyFast_Yes$Flight.Distance>=461 & survey_FlyFast_Yes$Flight.Distance<=848] = 'Medium'
survey_FlyFast_Yes$Flight.Distance.1[survey_FlyFast_Yes$Flight.Distance>=849 & survey_FlyFast_Yes$Flight.Distance<=1236] = 'Long'
survey_FlyFast_Yes$Flight.Distance <- NULL
View(survey_FlyFast_Yes)
survey_FlyFast_Yes$Flight.Distance.1 <- as.factor(survey_FlyFast_Yes$Flight.Distance.1)
```

By using the minimum and maximum values of flight distances which were 73 and 1236, we created the categories. Then dividing numbers into 3 categories. Flights of distance between 73 and 460 are short, 461 to 848 as Medium and 849 to 1236 as Long. After converting this newly created column to a factor we removed the old column Flight.Distance from the subset.

Creating Day.Of.Month.1 into category:

```
survey_FlyFast_Yes$Day.of.Month.1[survey_FlyFast_Yes$Day.of.Month>= 1 & survey_FlyFast_Yes$Day.of.Month <= 10] = 'Start month'
survey_FlyFast_Yes$Day.of.Month.1[survey_FlyFast_Yes$Day.of.Month>= 11 & survey_FlyFast_Yes$Day.of.Month <= 20] = 'Mid month'
survey_FlyFast_Yes$Day.of.Month.1[survey_FlyFast_Yes$Day.of.Month>= 21 & survey_FlyFast_Yes$Day.of.Month <= 30] = 'End month'
survey_FlyFast_Yes$Day.of.Month<- NULL
survey_FlyFast_Yes$Day.of.Month.1 <- as.factor(survey_FlyFast_Yes$Day.of.Month.1)
View(survey_FlyFast_Yes)
```

Creating 3 categories for the Month column. 1 to 10 as Start month, 11 to 20 as Mid month and 21 to 30 as End month. After converting this newly created column to factor we removed the old column Day.Of.Month from the subset.

Creating Scheduled Departure Period into category:

```
survey_FlyFast_Yes$Scheduled.Departure.period[survey_FlyFast_Yes$Scheduled.Departure.Hour>=0 & survey_FlyFast_Yes$Scheduled.Departure.Hour<=10] = 'Morning'
survey_FlyFast_Yes$Scheduled.Departure.period[survey_FlyFast_Yes$Scheduled.Departure.Hour>=11 & survey_FlyFast_Yes$Scheduled.Departure.Hour<=14] = 'Afternoon'
survey_FlyFast_Yes$Scheduled.Departure.period[survey_FlyFast_Yes$Scheduled.Departure.Hour>=15 & survey_FlyFast_Yes$Scheduled.Departure.Hour<=19] = 'Evening'
survey_FlyFast_Yes$Scheduled.Departure.period[survey_FlyFast_Yes$Scheduled.Departure.Hour>=20 & survey_FlyFast_Yes$Scheduled.Departure.Hour<=24] = 'Night'
survey_FlyFast_Yes$Scheduled.Departure.Hour <- NULL
survey_FlyFast_Yes$Scheduled.Departure.period <- as.factor(survey_FlyFast_Yes$Scheduled.Departure.period)
View(survey_FlyFast_Yes)
```


Dividing the Schedule departure hour into 4 categories, Hours 0 to 10 as Morning, 11 - 14 as afternoon, 15 - 19 evening and 20 - 24 as night. After converting this newly created column to factor and removing the old column Scheduled.Departure.Hour from the subset.

Creating Likelihood.to.recommend.1 into category:

```
survey_FlyFast_YesLikelihood.to.recommend.1[survey_FlyFast_YesLikelihood.to.recommend>=0 & survey_FlyFast_YesLikelihood.to.recommend<=4] = 'Low'
survey_FlyFast_YesLikelihood.to.recommend.1[survey_FlyFast_YesLikelihood.to.recommend>=5 & survey_FlyFast_YesLikelihood.to.recommend<=7] = 'Medium'
survey_FlyFast_YesLikelihood.to.recommend.1[survey_FlyFast_YesLikelihood.to.recommend>=8 & survey_FlyFast_YesLikelihood.to.recommend<=10] = 'High'
survey_FlyFast_YesLikelihood.to.recommend <- NULL
survey_FlyFast_YesLikelihood.to.recommend.1 <- as.factor(survey_FlyFast_YesLikelihood.to.recommend.1)
view(survey_FlyFast_Yes)
```

0 - 4 Recommendations are considered as Low. 5 - 7 as Medium and 8 - 10 as High.

After creating this column to factor we removed the old column Likelihood.to.Recommend from the subset.

Creating Rules using Apriori

Then using arules and arulesviz library, we found the rules using Apriori algorithm. The code and results from these rules can be seen below.

```
103 survey_FlyFast_NoX <- as(survey_FlyFast_No, "transactions") #Coercing the survey_FlyFast_No dataset into transaction matrix
104 survey_FlyFast_NoX
105 inspect(survey_FlyFast_NoX) #Using inspect function to display associations and transactions
106 rules <- apriori(survey_FlyFast_NoX, parameter = list(supp = 0.13, conf = 0.15),
107               control=list(verbose=F),
108               appearance=list(default="lhs",rhs=("Flight.cancelled=Yes"))) #Generating most relevant set of rules from given
109
110
111 inspect(rules[1:15])
112 #Texas, Georgia, Illinois are the origin state with 1 as confidence and lift from where most number of flights are getting cancelled
113 # Flights going to Illinois are mostly getting cancelled. (Confidence 1 and lift 1)
114
```

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{}	=> {Flight.cancelled=Yes}	1.0000000	1	1.0000000	1	425
[2]	{Origin.State=Texas}	=> {Flight.cancelled=Yes}	0.1317647	1	0.1317647	1	56
[3]	{Origin.State=Georgia}	=> {Flight.cancelled=Yes}	0.1317647	1	0.1317647	1	56
[4]	{Origin.State=Illinois}	=> {Flight.cancelled=Yes}	0.1341176	1	0.1341176	1	57
[5]	{Destination.State=Illinois}	=> {Flight.cancelled=Yes}	0.1435294	1	0.1435294	1	61
[6]	{Likelihood.to.recommend.1=Low}	=> {Flight.cancelled=Yes}	0.2094118	1	0.2094118	1	89
[7]	{Scheduled.Departure.period=Afternoon}	=> {Flight.cancelled=Yes}	0.2094118	1	0.2094118	1	89
[8]	{Day.of.Month.1=Mid month}	=> {Flight.cancelled=Yes}	0.2282353	1	0.2282353	1	97
[9]	{Day.of.Month.1=End month}	=> {Flight.cancelled=Yes}	0.2729412	1	0.2729412	1	116
[10]	{Flight.Distance.1=Medium}	=> {Flight.cancelled=Yes}	0.3223529	1	0.3223529	1	137
[11]	{Scheduled.Departure.period=Evening}	=> {Flight.cancelled=Yes}	0.3388235	1	0.3388235	1	144

With Support as 0.13 and confidence as 0.15, we found that with Confidence 1 and lift 1, Texas,

Georgia and Illinois are the destination states from where the flight originating are cancelled with 0.132, 0.132 and 0.134 support. With Support as 0.13 and confidence as 0.15, we found that with Confidence 1 and lift 1, Texas, Georgia and Illinois are the destination states from where the flight originating are cancelled with 0.132, 0.132 and 0.134 support. Illinois is the destination state that means, the flights going to Illinois are getting cancelled with Confidence 1 and lift 1 and support 0.143.

To find the impact of other variables, a new rule was created with Conf = 0.3 and support = 0.3. The top 3 rules are Airline status = Blue (with Confidence 1, Lift 1 and Support as 0.783 which is highest), Flight.Distance.1 = Short (with Confidence 1, Lift 1 and support 0.562 which is 2nd highest), and Day of month = Start month (with confidence 1, Lift 1 and support 0.498 which is 3rd highest). The code and results can be seen below:

```

115 rules1 <- apriori(survey_FlyFast_NoX, parameter = list(supp = 0.3, conf = 0.3),
116                   control=list(verbose=F),
117                   appearance=list(default="lhs",rhs=("Flight.cancelled=Yes"))) #Generating most relevant set of rules from giv
118
119
120 inspect(rules1)
121 # 49.88% of flights in the start of month is cancelled
122 # 56.23% of short distance flights are cancelled
123 # 78% of Blue airline status flights are cancelled
124
125

```

	lhs	rhs	support	confidence	coverage	lift	count
[1]	{}	=> {Flight.cancelled=Yes}	1.0000000	1	1.0000000	1	425
[2]	{Flight.Distance.1=Medium}	=> {Flight.cancelled=Yes}	0.3223529	1	0.3223529	1	137
[3]	{Scheduled.Departure.period=Evening}	=> {Flight.cancelled=Yes}	0.3388235	1	0.3388235	1	144
[4]	{Likelihood.to.recommend.1=High}	=> {Flight.cancelled=Yes}	0.3435294	1	0.3435294	1	146
[5]	{Scheduled.Departure.period=Morning}	=> {Flight.cancelled=Yes}	0.3600000	1	0.3600000	1	153
[6]	{Likelihood.to.recommend.1=Medium}	=> {Flight.cancelled=Yes}	0.4470588	1	0.4470588	1	190
[7]	{Day.of.Month.1=Start month}	=> {Flight.cancelled=Yes}	0.4988235	1	0.4988235	1	212
[8]	{Flight.Distance.1=Short}	=> {Flight.cancelled=Yes}	0.5623529	1	0.5623529	1	239
[9]	{Airline.Status=Blue}	=> {Flight.cancelled=Yes}	0.7835294	1	0.7835294	1	333
[10]	{Airline.Status=Blue,Likelihood.to.recommend.1=Medium}	=> {Flight.cancelled=Yes}	0.3811765	1	0.3811765	1	162
[11]	{Airline.Status=Blue,Day.of.Month.1=Start month}	=> {Flight.cancelled=Yes}	0.4023529	1	0.4023529	1	171
[12]	{Airline.Status=Blue,Flight.Distance.1=Short}	=> {Flight.cancelled=Yes}	0.4494118	1	0.4494118	1	191

2. Rules for likelihood of recommendation for Flyfast Airways.

We repeated this process to understand what influences the scores for FlyFast Airways.

We first created the subset of the dataframe. Before that, we again changed the numerical value

of Likelihood.To.Recommend column into the category of Low (Likelihood = 1 to 4), Medium (Likelihood = 5 - 7) and High (Likelihood = 8 - 10).

Creating the visualization to find which partnered airline is facing the most low Likelihood to recommend ratings. For this we are finding the rate of Low recommendation for each partnered airline. The formula is :

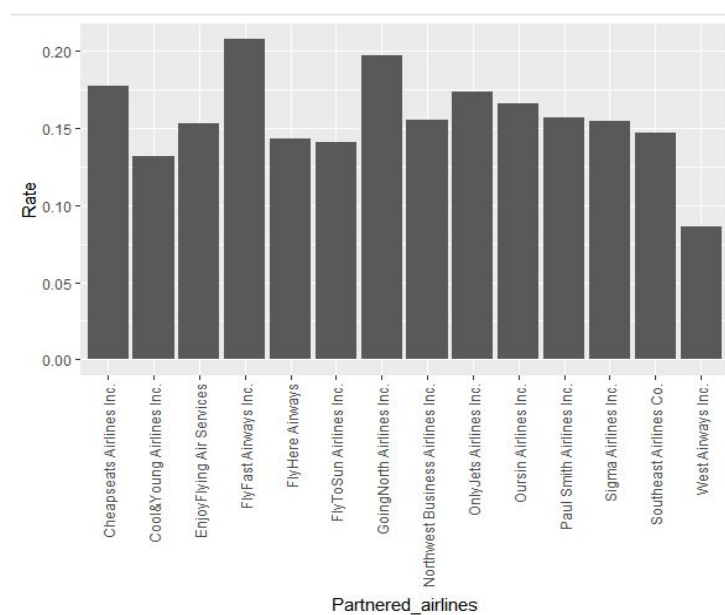
```
survey_cancelled$Rate <- survey_cancelled$Low / (survey_cancelled$High +  
survey_cancelled$Low + survey_cancelled$Medium)
```

Survey_Cancelled is the subset of the dataframe. The code snippet for the same is:

```
table(survey$Partner.Name, survey$Likelihood.to.recommend.1)  
survey_cancelled <- as.data.frame.matrix(table(survey$Partner.Name, survey$Likelihood.to.recommend.1))  
survey_cancelled$Rate <- survey_cancelled$Low / (survey_cancelled$High + survey_cancelled$Low + survey_cancelled$Medium)  
view(survey_cancelled)  
survey_cancelled$Partnered_airlines <- rownames(survey_cancelled)
```

Based on this newly created column, a bar graph is plotted. Ggplot2 is used for the visualization. Following is the code snippet of the same:

```
#Plotting the graph for rate of low recommended partnered airlines  
ggplot(survey_cancelled, aes(y=Rate, x=Partnered_airlines)) +  
  geom_bar(stat="identity") + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



We can clearly see that for Flyfast Airways, the rate of low recommendation is greater than 0.20 that is more than 20% of the recommendation are low which is maximum. We also notice West Airways is one of the better performing airlines with the lowest rate.

Creating the subset and manipulating the subset

Creating the survey_FlyFast dataset using below R code.

```
survey_FlyFast <- survey[(survey$Partner.Name == 'FlyFast Airways Inc.'],]
```

We Identified 10 independent variables which are Airline.Status, Age, gender, Class, Day.of.Month, Origin.State, Destination.State, Schedule.Departure.Hour, Flight.Time.in.minutes, Flight.Distance and Likelihood.To.Recommend as dependent variable.

Missing values

Replacing the missing or NA values in Flight.Time.in.minutes column with 0. Following is the R code used to do so:

```
# Converting flight.time.in.minutes column NA values to 0  
NA_Vec <- which(is.na(survey_FlyFast$Flight.time.in.minutes))  
survey_FlyFast[NA_Vec,10] <- 0
```

We also checked if there were any other missing or NA values in other variables and if they were present we omitted those rows. Following is the R code used for the same:

```
#Checking if there is a NA data and if present then removing them  
which(is.na(survey_FlyFast_Assoc))  
survey_FlyFast_Assoc <-  
survey_FlyFast_Assoc[complete.cases(survey_FlyFast_Assoc),]
```

##Creating Copy of dataset to survey_FlyFast_Assoc:

```
survey_FlyFast_Assoc <- survey_FlyFast
```

Converting the numeric data into categorical dataset

Flight.Distance:

As mentioned earlier, the three categories for this variable were:

Flight distance:73 - 460 - Short

Flight distance :461 - 848 - Medium

Flight distance :849 - 1290 - Long

```
#Scale: Flight distance :73 - 460 - Short
#Scale: Flight distance :461 - 848 - Medium
#Scale: Flight distance :849 - 1290 - Long

max(survey_FlyFast_Assoc$Flight.Distance)
min(survey_FlyFast_Assoc$Flight.Distance)
survey_FlyFast_Assoc$Flight.Distance.1[survey_FlyFast_Assoc$Flight.Distance>=73 & survey_FlyFast_Assoc$Flight.Distance<=460] = 'Short'
survey_FlyFast_Assoc$Flight.Distance.1[survey_FlyFast_Assoc$Flight.Distance>=461 & survey_FlyFast_Assoc$Flight.Distance<=848] = 'Medium'
survey_FlyFast_Assoc$Flight.Distance.1[survey_FlyFast_Assoc$Flight.Distance>=849 & survey_FlyFast_Assoc$Flight.Distance<=1290] = 'Long'
survey_FlyFast_Assoc$Flight.Distance <- NULL
View(survey_FlyFast_Assoc)
survey_FlyFast_Assoc$Flight.Distance.1 <- as.factor(survey_FlyFast_Assoc$Flight.Distance.1)
```

Day.of.Month:

As mentioned earlier, the three categories for this variable were:

Day of month :1 - 10 - Start month

Day of month :11 - 20 - Mid month

Day of month :21 - 30 - End month

```
#Scale: Day of month :1 - 10 - Start month
#Scale: Day of month :11 - 20 - Mid month
#Scale: Day of month :21 - 30 - End month

survey_FlyFast_Assoc$Day.of.Month.1[survey_FlyFast_Assoc$Day.of.Month>= 1 & survey_FlyFast_Assoc$Day.of.Month <= 10] = 'Start month'
survey_FlyFast_Assoc$Day.of.Month.1[survey_FlyFast_Assoc$Day.of.Month>= 11 & survey_FlyFast_Assoc$Day.of.Month <= 20] = 'Mid month'
survey_FlyFast_Assoc$Day.of.Month.1[survey_FlyFast_Assoc$Day.of.Month>= 21 & survey_FlyFast_Assoc$Day.of.Month <= 30] = 'End month'
survey_FlyFast_Assoc$Day.of.Month<- NULL
survey_FlyFast_Assoc$Day.of.Month.1 <- as.factor(survey_FlyFast_Assoc$Day.of.Month.1)
View(survey_FlyFast_Assoc)
```

Schedule.Departure.Period:

As mentioned earlier, the four categories for this variable were::

Flight Scheduled Departure Hour :0 - 10 - Morning

Flight Scheduled Departure :11 - 14 - Afternoon

Flight Scheduled Departure :15 - 19 - Evening

Flight Scheduled Departure :20 - 24 - Night

```
#Scale: Flight Scheduled Departure Hour :0 - 10 - Morning
#Scale: Flight Scheduled Departure :11 - 14 - Afternoon
#Scale: Flight Scheduled Departure :15 - 19 - Evening
#Scale: Flight Scheduled Departure :20 - 24 - Night

survey_FlyFast_Assoc$Scheduled.Departure.period[survey_FlyFast_Assoc$Scheduled.Departure.Hour>=0
& survey_FlyFast_Assoc$Scheduled.Departure.Hour<=10] = 'Morning'
survey_FlyFast_Assoc$Scheduled.Departure.period[survey_FlyFast_Assoc$Scheduled.Departure.Hour>=11
& survey_FlyFast_Assoc$Scheduled.Departure.Hour<=14] = 'Afternoon'
survey_FlyFast_Assoc$Scheduled.Departure.period[survey_FlyFast_Assoc$Scheduled.Departure.Hour>=15
& survey_FlyFast_Assoc$Scheduled.Departure.Hour<=19] = 'Evening'
survey_FlyFast_Assoc$Scheduled.Departure.period[survey_FlyFast_Assoc$Scheduled.Departure.Hour>=20
& survey_FlyFast_Assoc$Scheduled.Departure.Hour<=24] = 'Night'
survey_FlyFast_Assoc$Scheduled.Departure.Hour <- NULL
survey_FlyFast_Assoc$Scheduled.Departure.period <- as.factor(survey_FlyFast_Assoc$Scheduled.Departure.period)
view(survey_FlyFast_Assoc)
```

Age:

By using the minimum and maximum values of the Age which were 15 and 85, we created the categories. We divided the variable into 3 categories. These categories can be seen below:

Age :15 - 25 - Young

Age :26 - 50 - Adult

Age :51 - 85 - Old

```
#Scale: Age :15 - 25 - Young
#Scale: Age :26 - 50 - Adult
#Scale: Age :51 - 85 - Old

min(survey_FlyFast_Assoc$Age)
max(survey_FlyFast_Assoc$Age)
survey_FlyFast_Assoc$Age.1[survey_FlyFast_Assoc$Age>=15 & survey_FlyFast_Assoc$Age<=25] = 'Young'
survey_FlyFast_Assoc$Age.1[survey_FlyFast_Assoc$Age>=26 & survey_FlyFast_Assoc$Age<=50] = 'Adults'
survey_FlyFast_Assoc$Age.1[survey_FlyFast_Assoc$Age>=51 & survey_FlyFast_Assoc$Age<=85] = 'Old'
survey_FlyFast_Assoc$Age <- NULL
view(survey_FlyFast_Assoc)
survey_FlyFast_Assoc$Age.1<- as.factor(survey_FlyFast_Assoc$Age.1)
```

Flight.Time.in.Minutes:

By using the minimum and maximum values of the Flight Time which were 13 and 223, we created the categories. We divided the variable into 3 categories. These categories can be seen below:

Flight time: 13 - 90 - Short

Flight time: 91 - 150 - Medium

Flight time: 151 - 223 - Long

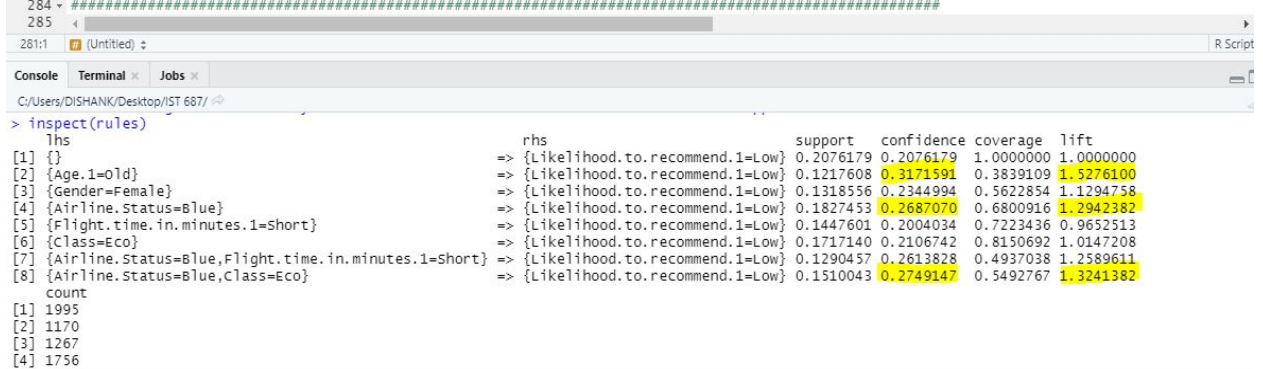
```
#Scale: Flight time :13 - 90 - Short
#Scale: Flight time :91 - 150 - Medium
#Scale: Flight time :151 - 223 - Long

max(survey_FlyFast_Assoc$Flight.time.in.minutes)
min(survey_FlyFast_Assoc$Flight.time.in.minutes)
survey_FlyFast_Assoc$Flight.time.in.minutes.1[survey_FlyFast_Assoc$Flight.time.in.minutes>=13
& survey_FlyFast_Assoc$Flight.time.in.minutes<=90] = 'Short'
survey_FlyFast_Assoc$Flight.time.in.minutes.1[survey_FlyFast_Assoc$Flight.time.in.minutes>=91
& survey_FlyFast_Assoc$Flight.time.in.minutes<=150] = 'Medium'
survey_FlyFast_Assoc$Flight.time.in.minutes.1[survey_FlyFast_Assoc$Flight.time.in.minutes>=151
& survey_FlyFast_Assoc$Flight.time.in.minutes<=223] = 'Long'
survey_FlyFast_Assoc$Flight.time.in.minutes <- NULL
view(survey_FlyFast_Assoc)
```

Apriori Rules

Using arules and arulesviz library, apriori algorithm is implemented to find the rules for Likelihood.To.Recommend for FlyFast Airways. The dataset is converted to a transaction table and then using support as 0.12 and confidence as 0.1, 8 rules are created. The code snippet is as follows:

```
271
272 #Creating Rules using Apriori
273 survey_FlyFast_Assocx <- as(survey_FlyFast_Assoc,"transactions") #Coercing the survey_FlyFast_No dataset into transaction matrix
274 survey_FlyFast_Assocx
275 inspect(survey_FlyFast_Assocx) #Using inspect function to display associations and transactions
276 rules <- apriori(survey_FlyFast_Assocx, parameter = list(supp = 0.12, conf = 0.1),
277               control=list(verbose=F),
278               appearance=list(default="lhs",rhs=("Likelihood.to.recommend.1=Low"))) #Generating most relevant set of rules from given tr
279
280 inspect(rules)
281 # Gender Female
282 # Status = Blue and Class Economy has highest Lift and coverage
283
284 #####
285
```



	lhs	rhs	support	confidence	coverage	lift
[1]	{}	=> {Likelihood.to.recommend.1=Low}	0.2076179	0.2076179	1.0000000	1.0000000
[2]	{Age.1=Old}	=> {Likelihood.to.recommend.1=Low}	0.1217608	0.3171591	0.3839109	1.5276100
[3]	{Gender=Female}	=> {Likelihood.to.recommend.1=Low}	0.1318556	0.2344994	0.5622854	1.1294758
[4]	{Airline.Status=Blue}	=> {Likelihood.to.recommend.1=Low}	0.1827453	0.2687070	0.6800916	1.2942382
[5]	{Flight.time.in.minutes.1=Short}	=> {Likelihood.to.recommend.1=Low}	0.1447601	0.2004034	0.7223436	0.9652513
[6]	{Class=Eco}	=> {Likelihood.to.recommend.1=Low}	0.1717140	0.2106742	0.8150692	1.0147208
[7]	{Airline.Status=Blue,Flight.time.in.minutes.1=Short}	=> {Likelihood.to.recommend.1=Low}	0.1290457	0.2613828	0.4937038	1.2589611
[8]	{Airline.Status=Blue,class=Eco}	=> {Likelihood.to.recommend.1=Low}	0.1510043	0.2749147	0.5492767	1.3241382

count

	count
[1]	1995
[2]	1170
[3]	1267
[4]	1756

The top 3 rules for low recommendation score that were identified are Gender = Female (Lift = 1.528 and Confidence = 0.317 which is highest), Airline status = Blue and Class = Eco (Lift = 1.32 and confidence - 0.274 which is 2nd highest) and Airline.Status = Blue (Lift = 1.294 and Confidence = 0.269 which is 3rd highest with support 0.121, 0.151 and 0.182 respectively.

Support Vector Machines

Our next modelling technique included Support Vector Machines. We developed an SVM data model to predict the Likelihood.To.Recommend for FlyFast Airways. As we have seen in previous association rules, FlyFast has the greatest number of Low recommendations. Hence we decided to create the model to predict the likelihood of recommendations by customers which will help them to predict whether customers will recommend them or not.

Using survey_FlyFast dataset that was created initially. The same 10 independent column were identified as Airline.Status, Age, Gender, Class, Day.of.Month, Origin.State, Destination.State, Scheduled.Departure.Hour, Flight.Time.in.Minutes, Flight.Distance and Likelihood.To.Recommend as dependent variable.

Converting non numeric columns to numeric

Converting columns like gender, Airline.Status and class which are non numeric to numeric columns.

```
#Converting string columns into numeric
survey_FlyFast$Airline.Status <- as.numeric(survey_FlyFast$Airline.Status)
survey_FlyFast$Gender <- as.numeric(survey_FlyFast$Gender)
survey_FlyFast$Class <- as.numeric(survey_FlyFast$Class)
str(survey_FlyFast)
view(survey_FlyFast)
```

Converting each column to factor:

The code used to convert each column can be to a factor can be seen below:

```
#Converting the variables into factor
survey_FlyFast$Airline.Status <- as.factor(survey_FlyFast$Airline.Status)
survey_FlyFast$Age <- as.factor(survey_FlyFast$Age)
survey_FlyFast$Gender <- as.factor(survey_FlyFast$Gender)
survey_FlyFast$Class <- as.factor(survey_FlyFast$Class)
survey_FlyFast$Day.of.Month <- as.factor(survey_FlyFast$Day.of.Month)
survey_FlyFast$Scheduled.Departure.Hour <- as.factor(survey_FlyFast$Scheduled.Departure.Hour)
survey_FlyFast$Flight.time.in.minutes <- as.factor(survey_FlyFast$Flight.time.in.minutes)
survey_FlyFast$Flight.Distance <- as.factor(survey_FlyFast$Flight.Distance)
survey_FlyFast$Likelihood.to.recommend.1 <- as.factor(survey_FlyFast$Likelihood.to.recommend.1)
```

Splitting the Dataset

Using Kernlab library, the dataset is split into a test set and a train set with a 60% split.

This meant the train dataset consisted of 60% of the rows which would be 5766 rows and the remaining 40% (3843 rows) would make up the test set. The code to do so can be seen below:

```
#Splitting data into training and testing dataset
library(kernlab) #Loading package kernlab
library(caret) #Loading package caret
trainList <- createDataPartition(y=survey_FlyFast$Likelihood.to.recommend.1,p=.60,list=FALSE) #Partitioning the 40% of data
trainset <- survey_FlyFast [trainList,] #Creating the training dataset using trainList
testSet <- survey_FlyFast[-trainList, ] #Creating test dataset using trainList
```

Creating and training SVM model

Likelihood.To.Recommendation will be the dependent variable as we are building a model to predict the Likelihood.To.Recommend for Flyfast Airways customers. For the model we are keeping $c = 5$. After training, Cross-validation came as 0.397 or approximately 40%.

Training error came out to be 0.052 which is approximately 5.2%

```
321
322 #Creating and training the svm data model with trainset
323 SVMoutput <- ksvm(Likelihood.to.recommend.1~., data = trainset, kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3, prob.model = TRUE)
324 SVMoutput
325 #Cross validation error 0.397]
326
327
325:30 (Untitled) R Script

Console Terminal Jobs
C:/Users/DISHANK/Desktop/IST 687/
Support Vector Machine object of class "ksvm"

SV type: C-svc (classification)
parameter : cost C = 5

Gaussian Radial Basis kernel function.
Hyperparameter : sigma = 0.0678733031674208

Number of Support Vectors : 4922

Objective Function value : -3453.729 -7188.217 -4420.012
Training error : 0.052723
Cross validation error : 0.396635
Probability model included.
```

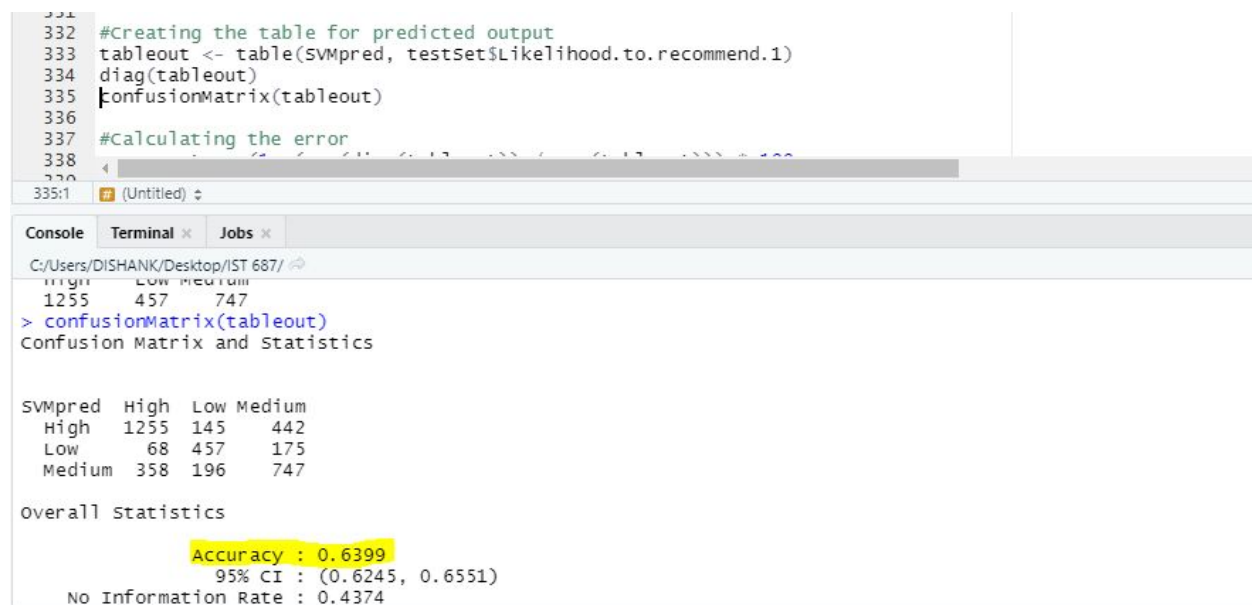

Predicting the Likelihood.To.Recommend with Testset

```
# Predicting the data model for testset
SVMpred <- predict(SVMoutput, newdata = testSet) #Predicting the class using test dataset
str(SVMpred) #viewing the structure of SVMpred
head(SVMpred) #viewing the first 6 rows of SVMpred
```

Creating a Confusion Matrix for the Model

From the Confusion Matrix, it is visible that Accuracy for the designed data model is 63.99% which is approximately **64%**. The matrix can be seen below:

```
331
332 #Creating the table for predicted output
333 tableout <- table(SVMpred, testSet$Likelihood.to.recommend.1)
334 diag(tableout)
335 confusionMatrix(tableout)
336
337 #Calculating the error
338
```



The screenshot shows the R console output for the confusion matrix and overall statistics. The confusion matrix is as follows:

SVMpred \ Likelihood	High	Low	Medium
High	1255	145	442
Low	68	457	175
Medium	358	196	747

Overall Statistics:

- Accuracy : 0.6399
- 95% CI : (0.6245, 0.6551)
- No Information Rate : 0.4374

1255 High recommended values are correctly identified as High but 68 and 358 number of values were incorrectly predicted as Low and Medium respectively instead of High. 457 Low recommended values were correctly identified as Low. But 145 and 196 values were incorrectly predicted as High and Medium respectively. 747 values were correctly predicted as Medium. But 442 and 175 values were incorrectly predicted as High and Low respectively.

Calculate the percentage error for designed SVM Model

Percent error is **36.01%** which is shown in the below screenshot.

```
balanced Accuracy      0.7373      0.7404      0.6621
> #Calculating the error
> error_rate = (1 -(sum(diag(tableout)) / sum(tableout))) * 100
> error_rate
[1] 36.01353
>
```

Linear Regressions

In this section we will be talking about the different types of linear regressions we used to model to determine the factors that influence a higher likelihood to recommend score. As we already conducted an in-depth analysis regarding FlyFast Airways, the worst performing airline, we used these linear models to conduct a more holistic analysis. The findings from these models did further show us that FlyFast Airways is the worst performing airline. We ran a total of three linear regressions, with each model's dependent variable being "Likelihood.to.recommend". The three models attacked what we believe are the three major aspects that can help understand what can improve the score. Within these models we used a variety of dummy variables. As mentioned earlier a dummy variable is simply a binary column of 1 or 0, with 1 being if the condition is true and 0 if false. For example the dummy variable for Gender_Male would be a 1 if the customer is male and 0 if the customer is not male. The three models are:

1. Flight and Airport Factors

For this linear model we used different variables that related to the flight and the airport as the independent variables. The independent variables we used included:

- a. Price Sensitivity
- b. Shopping Amount
- c. Eating and Drinking Amount
- d. Flight Time

- e. Delay in Arrival
- f. Scheduled Departure Hour
- g. Flight Cancellation

The idea between this regression was to understand what flight types tend to have a positive impact on consumers. Below is a screenshot of the model's output in R:

```
Call:
lm(formula = Likelihood.to.recommend ~ -1 + Price.Sensitivity +
    Shopping.Amount.at.Airport + Eating.and.Drinking.at.Airport +
    Flight.time.in.minutes + Arrival.Delay.in.Minutes + Scheduled.Departure.Hour +
    Flight.cancelled, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-12.7574  -1.4172   0.8079   2.4694   7.8217

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
Price.Sensitivity    1.4103532   0.0144831   97.38  <2e-16 ***
Shopping.Amount.at.Airport 0.0043112 0.0001765   24.43  <2e-16 ***
Eating.and.Drinking.at.Airport 0.0138017 0.0001721   80.19  <2e-16 ***
Flight.time.in.minutes    0.0091766 0.0001235   74.29  <2e-16 ***
Arrival.Delay.in.Minutes  -0.0040950 0.0002497  -16.40  <2e-16 ***
Scheduled.Departure.Hour    0.2374528 0.0015683  151.41  <2e-16 ***
Flight.cancelled          0.7606979 0.0714228   10.65  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.811 on 88093 degrees of freedom
Multiple R-squared:  0.8649,    Adjusted R-squared:  0.8648
F-statistic: 8.054e+04 on 7 and 88093 DF,  p-value: < 2.2e-16
```

The adjusted R-Squared for this model is 0.8648, this entails the model does a good job in predicting the score. All the variables are also statistically significant which can be seen by the P - Value for each variable. Based on the model we can see the most positive impact on a higher score is the Price Sensitivity. Flight tickets that have a higher price sensitivity tend to receive higher scores. The intuition behind this in our opinion is that if a ticket with a high price sensitivity is bought, it would mean a consumer is most likely buying the ticket due to a decrease in the price. Hence, the cheaper ticket provides them with more value resulting in a higher recommendation score. A surprising result is the Cancelled Flights tend to have a positive impact

on the recommendation score. We believe that this is the case because, consumers are adequately compensated for this and the customer service displayed by the airline is extremely well done.

As expected, flights that have a delay result in the Recommendation Scores being lower.

2. Consumer Factors

For this linear regression we used consumer factors. Essentially we want to understand what type of customers tend to provide higher recommendation scores as we can focus our marketing efforts towards those consumers. As per the box plot we discussed earlier in the paper, we expected that among the type of travel, Business reasons, would have the greatest positive impact on a recommendation score. In addition to this, we expected male passengers to provide higher scores from the boxplot. The linear model used a variety of independent variables that would provide us with far more insight than the boxplot. The independent variables we used included:

- a. Age
- b. Type of Travel - Business
- c. Type of Travel - Personal
- d. Type of Travel - Mileage Tickets
- e. Year of First Travel
- f. Flights per Year
- g. Loyalty
- h. Total Frequent Flyer Accounts
- i. Gender - Male

Below is a screenshot of the model's output in R:

```
Call:
lm(formula = Likelihood.to.recommend ~ -1 + Age + Year.of.First.Flight +
    Flights.Per.Year + Loyalty + Type.of.Travel + Total.Freq.Flyer.Accts +
    Gender_Male, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-8.4558 -1.2078  0.4335  1.4090  5.3234

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
Age           -2.197e-03  4.312e-04  -5.094 3.51e-07 ***
Year.of.First.Flight  1.001e-02  2.110e-03   4.742 2.12e-06 ***
Flights.Per.Year    -1.424e-02  6.412e-04 -22.215 < 2e-16 ***
Loyalty           -4.131e-02  1.749e-02  -2.362 0.018198 *
Type.of.TravelBusiness travel -1.167e+01  4.235e+00  -2.756 0.005849 **
Type.of.TravelMileage tickets -1.196e+01  4.235e+00  -2.825 0.004733 **
Type.of.TravelPersonal Travel -1.420e+01  4.235e+00  -3.354 0.000796 ***
Total.Freq.Flyer.Accts  -5.680e-03  6.414e-03  -0.886 0.375803
Gender_Male         1.784e-01  1.288e-02  13.854 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.863 on 88091 degrees of freedom
Multiple R-squared:  0.9406,    Adjusted R-squared:  0.9406
F-statistic: 1.55e+05 on 9 and 88091 DF,  p-value: < 2.2e-16
```

The adjusted R-Squared for this model 0.9406. This is extremely close to 1 which entails that this model is fairly accurate. Based on the results we can clearly see that younger consumers tend to provide higher recommendation scores. Age has a negative coefficient implying that the younger you are, the higher the recommendation score is to be. We also see that Year of First Flight has a positive coefficient which also implies a younger consumer group as they only began flying recently. In addition to this male consumers tend to provide greater recommendation scores. Lastly, we can see that among all three of the travel types, Business Travel has the largest coefficient implying that customers travelling for business reasons tend to provide higher recommendation scores in comparison to those customers travelling for personal reasons or using mileage tickets.

3. Airlines

This regression looked at each airline to understand which specific airlines usually have higher recommendation scores. To create this model, we created a dummy variable for each airline. The model's output can be seen below:

```
Call:
lm(formula = Likelihood.to.recommend ~ -1 + ., data = airline)

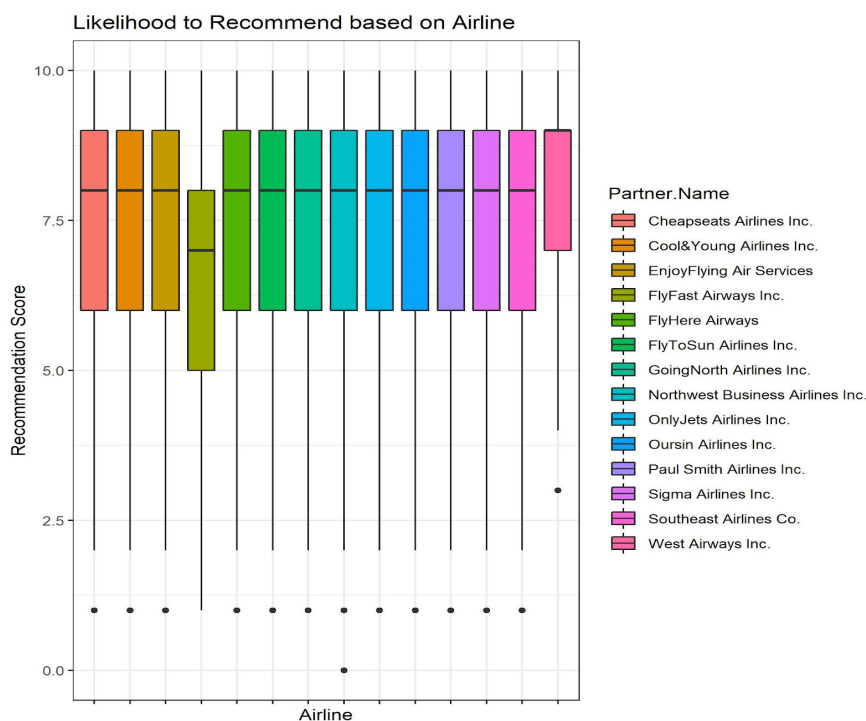
Residuals:
    Min       1Q   Median       3Q      Max
-7.4721 -1.4441  0.5314  1.6584  3.3352

Coefficients:
`Partner.Name_Cheapseats Airlines Inc.`      7.24708  0.01618  447.81  <2e-16 ***
`Partner.Name_Cool&Young Airlines Inc.`      7.58772  0.06962  108.98  <2e-16 ***
`Partner.Name_EnjoyFlying Air Services`      7.45297  0.03411  218.48  <2e-16 ***
`Partner.Name_FlyFast Airways Inc.`          6.66484  0.02224  299.67  <2e-16 ***
`Partner.Name_FlyHere Airways`              7.49705  0.04945  151.61  <2e-16 ***
`Partner.Name_FlyToSun Airlines Inc.`        7.54291  0.04425  170.46  <2e-16 ***
`Partner.Name_GoingNorth Airlines Inc.`      7.24084  0.06226  116.30  <2e-16 ***
`Partner.Name_Northwest Business Airlines Inc.` 7.47207  0.02156  346.64  <2e-16 ***
`Partner.Name_OnlyJets Airlines Inc.`        7.34165  0.04013  182.94  <2e-16 ***
`Partner.Name_Oursin Airlines Inc.`          7.28928  0.02497  291.89  <2e-16 ***
`Partner.Name_Paul Smith Airlines Inc.`      7.44413  0.03298  225.72  <2e-16 ***
`Partner.Name_Sigma Airlines Inc.`          7.46859  0.01907  391.62  <2e-16 ***
`Partner.Name_Southeast Airlines Co.`       7.46942  0.02531  295.17  <2e-16 ***
`Partner.Name_West Airways Inc.`            8.05172  0.20706   38.88  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.23 on 88086 degrees of freedom
Multiple R-squared:  0.9149,    Adjusted R-squared:  0.9149 
F-statistic: 6.766e+04 on 14 and 88086 DF,  p-value: < 2.2e-16
```

The model has a high adjusted R-Squared value of 0.9149. Based on the coefficients of the independent variables we can see that the airline that tends to receive lower scores are FlyFast

Airways. The highest is West Airways. We further explored these results with a boxplot as well to make it easier to understand and more visually appealing. On the left we can see a box plot depicting that most airlines are extremely similar with their recommendation



scores, the only two outliers being FlyFast Airways and West Airways. Considering West Airways' success it may be useful to understand their customer service training and plans and see if those strategies can be adapted into other airlines, particularly FlyFast Airways. As expected, this model and visualization simply solidifies our reasoning for focusing on FlyFast Airways and attempting to improve their performance.

Findings and Recommendations

Based on our modelling and findings, we have decided to break down our recommendations into two broad categories, similar to how we focused our analysis on 2 aspects. The first grouping of recommendations is specifically for FlyFast Airlines and how they can improve their Net Promoter Score, the second grouping of recommendations are for all airlines that are a part of the Southeast family. While these airlines perform well, we want to strive for the best and are hence suggesting possible ideas that can further improve their performance.

1. FlyFast Airline Findings, Recommendations, Solutions

As we noticed with our exploratory data analysis, FlyFast Airlines was one of the worst performing airlines in the dataset. They were the only airline with a negative Net Promoter Score and had the most flights cancelled among all of Southeast's regional partners. We used an association rules model to determine the reasons for low recommendation scores from customers and the reasons for flights being cancelled often for FlyFast Airways.

Based on our first association rules model, we narrowed down on 3 factors that lead to FlyFast Airways being cancelled:

1. Flights originating from Texas, Georgia, and Illinois
2. Flights with a flight distance that was considered short
3. Flights that occurred earlier in the month

While obviously it is not feasible to simply stop having flights originating from the states of Texas, Georgia, and Illinois, we do believe that it may be beneficial to reduce the number of flights originating from these locations. In addition to this, we would strongly recommend to either decrease the number of short flights offered or to invest more into the quality of airplanes

used for these shorter flights. Lastly we believe that flights occurring earlier in the month tend to be cancelled due to the lower demand during these flights. Our suggestion would be to either increase the price of these tickets as it would make the flights feasible even if there is a low number of passengers or to reduce the number of these flights.

We do want to make it clear we are not recommending to lower the total number of flights offered, we simply suggest that flights should be offered from different locations or at different times based on our finding.

Based on our second association rules model for FlyFast Airways we identified 3 major reasons for low recommendation scores:

1. Female Customers
2. Customers with an airline status of Blue
3. Customers travelling in Economy Class

Our main recommendation from this model would be to work on improving the services provided by FlyFast Airways towards new customers. Customers with a status of blue tend to be new customers and most new customers for an airline travel in economy class. Clearly the services provided by FlyFast airways are not up to the mark. We believe working on improving the experience of customers in the economy class will drastically improve the net promoter score. While it is fairly obvious, improving customer experience in an area will improve the performance of the airline. We believe that the marginal effect on improving customer experiences in the economy class is the largest in terms of impacting FlyFast Airway's net promoter score. This means that the increase in performance will be greatest if FlyFast Airways improves their economy class experience for customers.

Lastly, to offer a service that can help FlyFast Airways target customers that are more likely to provide a higher recommendation score, we created a support vector machine. This model has an accuracy of 64% and will predict if a customer is to provide a high, medium, or low recommendation score. This model can be used to understand which consumer should be focused on when marketing FlyFast Airways. Promotional material can be tailored to be sent to those customers that are predicted to provide higher recommendation scores.

2. Overall Findings and Recommendations based on Customer and Flight Attributes

To ensure we provide insight to every airline that is a part of the Southeast family, we also looked into general attributes for Customers and Flights that lead to higher recommendation scores. Using linear modelling we identified the following attributes tend to lead to higher recommendations:

1. Younger Customers
2. Male Customers
3. Customers Travelling for Business
4. Longer Flights
5. Flights Scheduled to Depart Later in the Day

As all of the linear models we used had an extremely adjusted R Squared value, we are confident that these attributes correlate with higher recommendation scores. Based on our modelling, we strongly believe there is an emerging market that airlines can benefit in the form of a new workforce. This entails those young adults who are currently just entering the workforce and have not flown much. This also connects to our initial exploratory data analysis where we saw higher recommendation scores from males travelling for business reasons. We recommend

airlines to have their marketing efforts focused towards a younger male audience looking to travel for business.

From a flight selection perspective, it may benefit airlines to look into increasing the number of flights that are scheduled later in the day (late evening or night time) and longer in nature (in terms of time). These flights tend to rub customers in a positive manner as they provide higher recommendation scores after these flights.

Conclusion

We strongly believe that our analysis and modelling have revealed key aspects that can help Southeast Airlines and their regional partners. We hope our focus on FlyFast Airways can help improve the performance of this airline, bring it up to the standard of the other airlines. We hope our overall recommendations strongly help in improving the overall Net Promoter Score for each airline as well.

In conclusion, our initial exploration and analysis of the data revealed key characteristics present among the airlines, particularly that FlyFast Airways was not performing as well as other airlines. Hence we focused a large amount of analysis on the issues with FlyFast Airways and how we could improve this airline. Using an association rules model and support vector machine we identified the best attributes to focus on improving FlyFast Airways. We also analysed factors that could benefit all airlines as there is always room for improvement. We did so by employing a linear regression that identified key characteristics of customers and flights that can help an airline improve their performance and net promoter score.