Southeast Airlines:
Combating Customer Churn

Final Project Report

By: Ryan Reed, Sidney Lanier, Abhiram Gopal, Renjie Zhu, Scott Kessel

IST 687 M002 Group 2

Table of Contents

Introduction (Scope/Context/Background/Business Questions)

This report outlines the requested analysis from Southeast Airlines who has asked our team (Group 2) to perform data analytical analysis to lower their customer churn or in simple terms – help reduce customer attrition. Up until today Southeast airlines has been dependent on a robust loyalty or frequent flyer program in order to reduce customer attrition. However, Southeast Airlines has grown curious as to whether this is truly the case due to recent studies, such as, the International Air Transport Association (IATA) study which stated that "airlines carry $12B in "loyalty debt" and frequent flier mileage and points are slowly devaluing, while the overall balance (or debt) is increasing." [1]

Therefore, Southeast Airlines has tasked Group 2 to study past records of their airline and partner airlines in order to proactively reduce churn by cutting off the head of the snake (so-to-speak) by getting in front of potential future customer loss by finding instances of significant indicators, or metrics, to aid in keeping customers. By identifying said metrics Southeast Airlines could use this knowledge to know when a customer was about to stop flying Southeast and try and prevent it. The analysis that will take place in this report will highlight suggestions as to how to avoid having the customers leave and go to another airline. In thinking about customer churn, several key facts are relevant:

1. As stated above the overall goal that Southeast Airlines wished for group 2 to provide actionable insight, based on the data available.
2. In order to provide the requested actionable insight based on the data available the business questions that must be answered for this project are:
    a. How do we improve the Net Promoter Score for Southeast Airlines?
    b. Which attributes from the data are most interesting when considering NPS?
    c. What are the traits of customers who are most likely to give detractor scores when asked about likelihood to recommend?
        i. Which group is most likely to give passive scores?
    d. Which attributes are most important to promotor, passive, and detractor customers and how do they impact the likelihood to recommend score?

Net Promoter Score (NPS)

NPS (Net promoter score) is a management tool used to gauge the loyalty of customer relationship. A NPS uses a 1-10 scale which represents the answer to the simple question: "How likely is it that you will recommend our airline to a friend or colleague?". This score in turn represents 3 categories of which customer falls into - promoters, detractors & passives. If respondents score is 1-6, it is considered detractors. If they score is 9-10, it is promoters, and if the score is in the middle range (a score 7-8), the score is "passive". In order to determine the

overall NPS score a subtraction of the percent of records who are classified as detractors from promoter's records needs to occur. The reasoning behind the importance of NPS is based on the fact that customers who are classified as promoters are considered loyal customers who may spread free "word of mouth" advertising to help stimulate further customer expansion. Oh the other hand customers who are classified as detractors are considered a threat and may spread negative advertisements through their social connections. (i.e., tell others not to fly Southeast Airlines). The strategy for group 2 is to find the factors responsible within the Southeast Airline data that makes their customers passive and not a promoter. Furthermore, what makes their customers fall into the detractor zone? Which will actually give us harmful reviews.

The Data Available

Southeast Airlines and their corresponding airline partners that operated within their company have taken great strides to gain feedback from their customers. This crucial feedback, of possessed recently completed customer survey data has been used by Southeast in order to calculate NPS. However, now that Southeast has hired Group 2 they will be providing our team with the data in order to analyze and increase their focus on providing good customer service when their NPS score went down.

Please note that the below description of the data provided to us is a direct description taken from the Southeast airline survey staff for the purposes of understanding the dataset that was provided. For the sake of clarity by both companies group 2 though is acceptable to take this excerpt directly from the original data description document provided.

"The survey dataset contained thousands of observations of flight segment data collected by Southeast Airlines. Each row represents one flight segment, by one airline (either southeast or one of its partner airlines), for a specific customer. Each column represents an attribute of that particular flight segment. Each row captures 26 characteristics of the flight (ex. day of month, date, airline, origin and destination city, if the flight was delayed), the customer (ex. age, gender, price sensitivity, the person's frequent flyer status). The row also contains a simple survey-based rating of each customer's likelihood to recommend the airline that they just flew as well as a field for open-ended text comments. It should be noted that there are some missing values in the dataset. The table below provides a short description for each attribute." [1]

Attributes:

1. ***Likelihood to Recommend –*** *rated on a scale of 1 to 10, which shows how likely the customer is to recommend the airline to their friends (10 is very likely, and 1 is not very likely).*

2. ***Airline Flyer Status*** *– each customer has a different type of airline status, which are platinum, gold, silver, and blue (based on level of travel with the airline)*
3. ***Age*** *– the specific customer's age. Ranging from 15 to 85 years old.*
4. ***Gender*** *– male or female.*
5. ***Price Sensitivity*** *– the grade to which the price affects to customers purchasing. The price sensitivity has a range from 0 to 5.*
6. ***Year of First Flight*** *– this attribute shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.*
7. ***Flights Per Year*** *– The number of flights that each customer has taken in the most recent 12 months. The range starting from 0 to 100.*
8. ***Loyalty*** *– An index of loyalty ranging from -1 to 1 that reflects the proportion of flights taken on other airlines versus flights taken on this airline. A higher index means more loyalty.*
9. ***Type of Travel*** *– One of business travel, mileage tickets, or personal travel (ex. vacation)*
10. ***Total Frequent Flyer Accounts*** *– How many frequent flyer accounts the customer has.*
11. ***Shopping Amount at Airport*** *– The spending on non-food & services at the airport (in $)*
12. ***Eating and Drinking at Airport*** *– The spending on food/drink at the airport (in $).*
13. ***Class*** *– three different kinds of service level (business, economy plus, and economy).*
14. ***Day of Month*** *–the traveling day of each costumer (ranges from 1 to 31).*
15. ***Flight date*** *– the passenger's flight date of travel.*
16. ***Partner Code*** *– This airline works with wholly- and partially-owned subsidiary companies to deliver regional flights. For example, AA, AS, B6, and DL.*
17. ***Partner Name*** *– These are the full names of the partner airline companies.*
18. ***Origin City*** *– the place where passenger departed from. For example, Boston MA.*
19. ***Origin State*** *– the place where passenger departed from. For example, Texas.*
20. ***Destination City*** *– the place to which passenger travels to. For example, Boston MA.*
21. ***Destination State*** *– the place to which passenger travels to. For example, Texas.*
22. ***Scheduled Departure Hour*** *– the specific time at which the plane was scheduled to depart.*
23. ***Departure Delay in Minutes*** *– How long the flight's departure was delayed, when compared to schedule.*
24. ***Arrival Delay in Minutes*** *– How long the arrival was delayed.*
25. ***Flight Cancelled*** *– occurs when the airline does not operate the flight.*
26. ***Flight time in minutes*** *–the length of time, in minutes, to reach the destination.*
27. ***Flight Distance*** *– the distance between the departure and arrival destination.*
28. ***Comment*** *– a free form text field of the passenger comment, with respect to the flight.*

Data Acquisition, Cleaning, Transformation, Munging

Before we even start any modelling it is important we make sure we clean the data first. Cleaning here means:

1. NA interpolation
2. Outlier analysis
3. Binning variables

In our work, we deep dive into each column, check for outliers, and NAs. We also check whether particular columns need binning. For some columns, we combine them together to create a new column - (Time delays). After the initial cleanup (during which the above 3 steps are done), we start with the basic mining to find initial trends. At the end of this section, we shall have the final code, which can be plugged directly into models below and our action plan (based upon the preliminary analysis) to select the types of models.

Data Acquisition and Cleaning

In this section, we load the data and discuss how we filled the NAs and also the outlier analysis of columns.

The dataset has 32 total columns. Each column was checked for for the following aspects and cleaned accordingly.

1. Character columns - Other than the FREE-TEXT column all other character type columns were converted into factors. All columns were made sure that the factors are not redundant(For example : Having 2 factors like ; ALABama and alabama)
2. Numeric columns - The majority of the columns were of numeric type. But it doesn't make sense to have columns having date , factor type values as numeric. So appropriate conversions are done. When dealing with the numeric columns, we checked into the box-plots of each column and did a suitable substitution of NAs . As for outliers, most outliers weren't far off from the 99th percentile and since our models mostly used binned variables for analysis, we did not do much of outlier modifications like “winsorizing or converting them to mean..etc”. Only four numerical columns had NAs.

NA Interpolation

For departure delay and arrival delay in minutes, we replaced the NAs with the mean of the columns because the mean was an accurate enough representation of the delay.

For the flight time in minutes variable our strategy was to replace the NAs with flight times that

have the same origin city and destination city from other records (rows) in the (i.e. the same flight just from a different record or person). By implementing this strategy we were able to find more realistic times for each NA for this variable instead of a generic replacement of the mean value of this variable. It was found however that the flights which have no existing flight time records ("Milwaukee, WI" to"Minneapolis, MN"), our group had to google maps the flight times and populated flight time for these NAs individually. Please note that this strategy was only ideal because there were few instances of this occurrence if there were multiple instance an automated search of google maps and population into the dataframe would need to be coded.

For the missing values in the likelihood to recommend column, we substituted the missing value with the mean of score with respect to the same airlines. for example there was a A in Northwest Business airlines. We interpolated that value of NA using the rounded mean of all likelihood to recommend scores of that particular airline.

Outlier Analysis

Checking for outliers is a crucial step in ensuring that our data is clean. In our modelling analysis we do not want to have models with outliers, which can lead to wrong results. For numerical variables the primary step of outlier analysis is to visualize the column using a box plot. (Refer Figure 1).

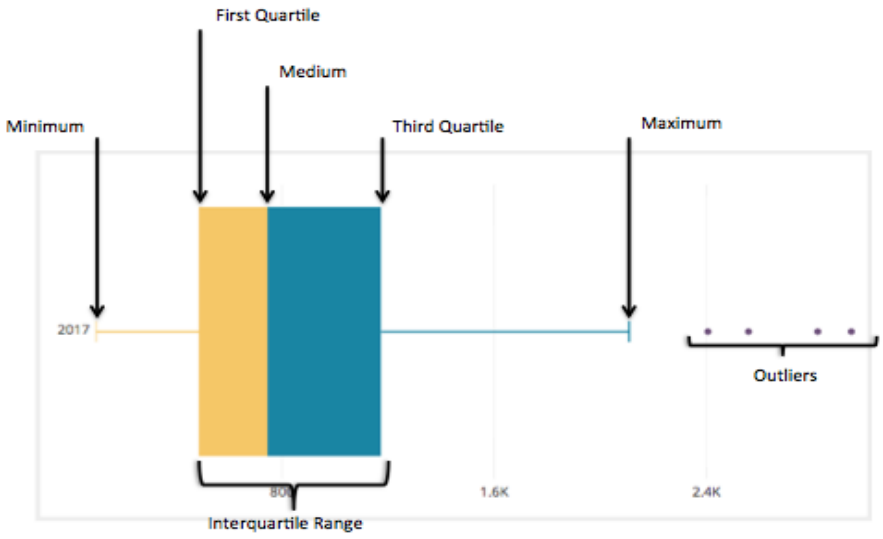

Figure 1 - Figure of a Box plot

Once we observe outliers, there are different methods to approach them.

1. Winsorize

2. Converting them any measures of central tendency
3. Interpolations
4. Binning them

In our project we went by method four, since our primary concern is analysis and not much of modelling. Binning gave us an easier to read and interpret sample. So for accurate results we felt it would be better to have 4-5 bins instead of a continuous range of numbers.

Binning & Categorizing

A total of 8 columns were binned. Bins were made from using the information from the metadata and from the column's numerical descriptive statistics(Like quantiles,mean,median..etc).
Also additional columns were created which we felt relevant to the analysis.

Descriptive Statistics & Visualizations

Before we go into any major models or advanced regressions, we should first try to understand our data through the help of simple visualization graphs by connecting variables. In this project as defined above our objective is to work on passives, detractors and promoters. Let us first understand what we have in our data.

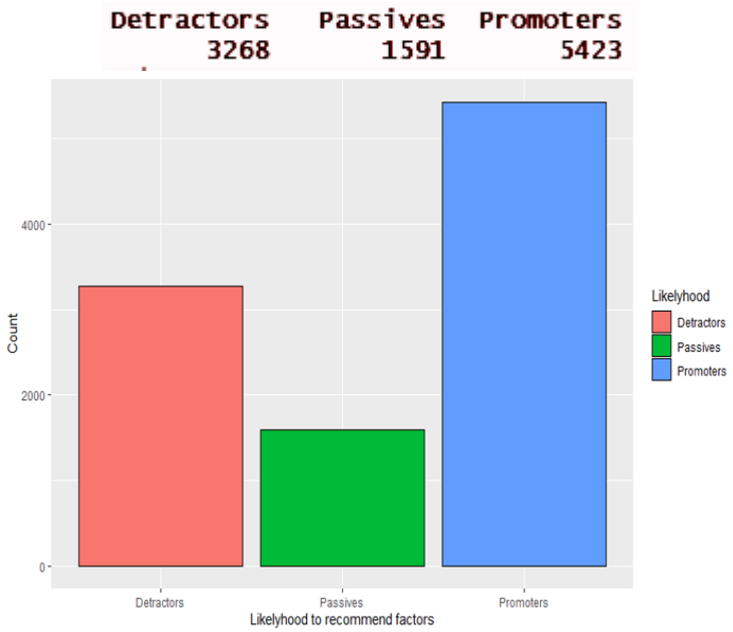

Figure 2 - number of detractors, promoters and passives

We can see from the above graph that we have 50% of our data as passives and detractors. Next we wanted to see as per airlines. There are 14 partner airlines in the entire dataset. Our goal is to

make sure that we select the best airlines as our dataset so that our resulting analysis yields the most statistically significant results.

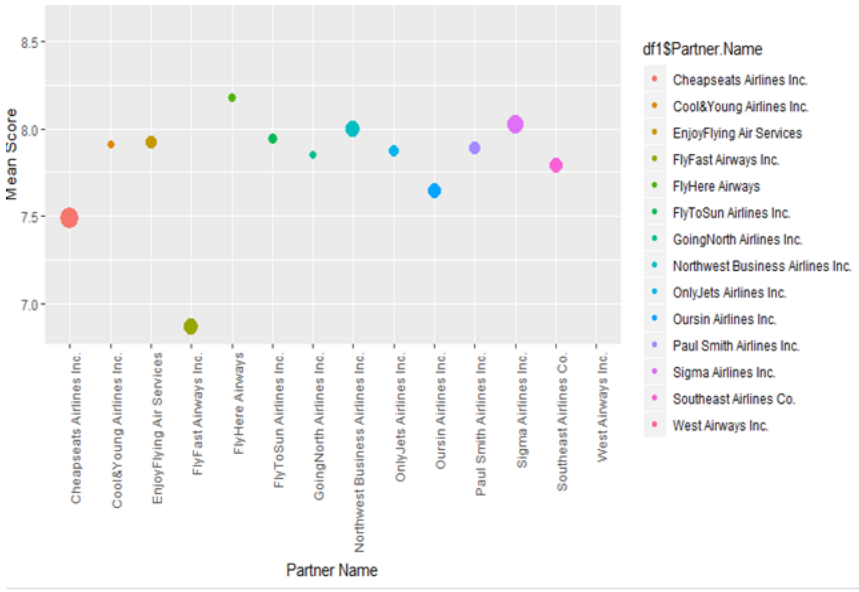

Figure 3 - Mean score of each airline (size of bubble represents its number of observation)

From the above we have 9 airlines in the passive region , 1 airline in detractor. But then our goal here is to select the “best sample” so that we can have better results and correlated models.(which compliment each other).

Determining Appropriate Dataframes to Analyze

In this section we are going to interpret the overall goal of this project and look at the overall data and variables we wish to increase to satisfy said overall goal. Please remember our overall goal is to increase the NPS score for the airline company. Through analysis of the airline data provided to us we have determined that the Likely.hood.to.recommend variable is the variable that we are to consider for calculating (and maximizing) the NPS score. Therefore, our approach is to focus and analyze data that is considered the most influential to maximizing the NPS score. Our group strategy to do this is as follows… First, we separated the large dataframe into 3 dataframes, the Detractor dataframe (likely.hood.to.recommend variable = 0-6), the Passives dataframe (likely.hood.to.recommend variable = 7-8), and the Promoters dataframe (likely.hood.to.recommend variable = 9-10).

Please note that at this point we analyzed the 3 dataframes and realized that in order to increase the NPS (i.e. NPS = Promoters - Detractors) we need to worry about increasing the Detractors scores into Passives or Promoters (thus decreasing our detractor value in the NPS equation) and increasing the Passive scores into Promoters (thus to increase the Promoter score in the NPS equation) which will then increase the airlines overall NPS. By implementing this strategy we

determined that the 3755 records that are in the promoter category are not reliant on our because the Promoters are already in the category we wish to have them in, therefore, the dfPromotor dataframe was neglected from any further analysis.

At this point an analysis of the 14 airlines sampling size based on Detractor and Passive values occurred in order to determine which of the 14 airlines had significant sampling sizes in order to influence the overall dataset (i.e. all 10282 records) thus in turn influencing our NPS. The first steps we took to determine this analysis was to plot two graphs. The first Figure 4: shows the distribution of Likely.hood.to.recommend variable (NPS value) Detractor Values by airline (Partner.Name variable). As one can see by viewing this graph the majority of the Detractor values (Likely.hood.to.recommend value of 0-6) are present in 6 airlines while the remaining 8 airlines have a smaller distribution of Detractor values.

The second graph that was created shows a similar distribution however this distribution is based on Passive values (Likely.hood.to.recommend variable values of 7-8). In Figure 5 one can see a distribution of Passive values with 6 airlines (the same 6 airlines from Figure 4) displaying the majority of the Passive values. This is important because we are attempting to prove that an analysis of all 14 airlines is not needed. The amount of the sample size in each airline is important towards influencing the overall population and thus the NPS result. Therefore, the higher the sum total of relevant records (in this case passive scores) within each airline indicates the better chance that said airline can influence (or raise) the total Passive value over other airlines. Additionally, by analyzing these more influential airlines customers can gain insight into significant variables or key customers they should focus on in order to raise the overall NPS rating. The key takeaway from the Figure 5 graph is that the same 6 airlines are relevant for Passive scores as were for Detractor scores in Figure 4.

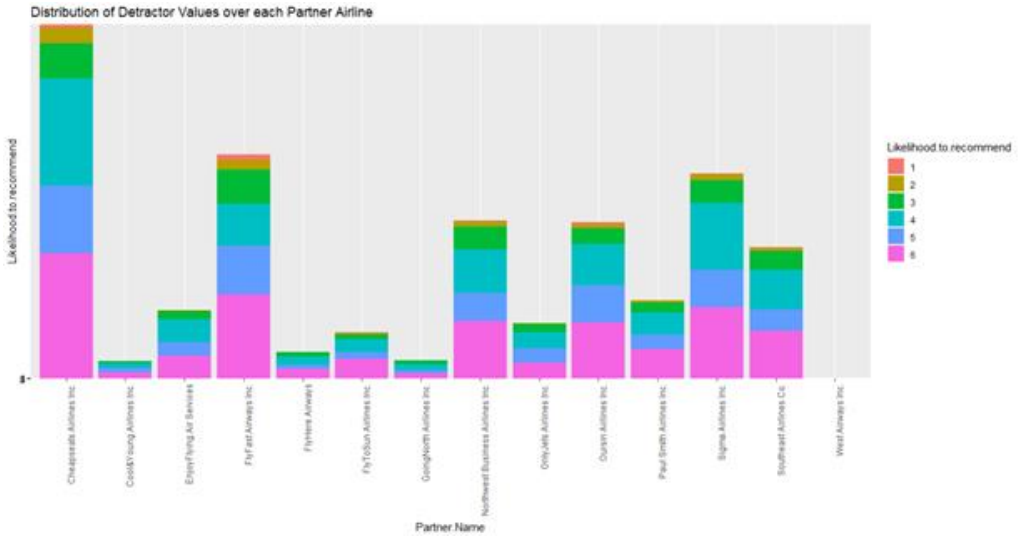

Figure 4 - Show the distribution of Likely.hood.to.recommend (NPS value) Detractor Values by airline

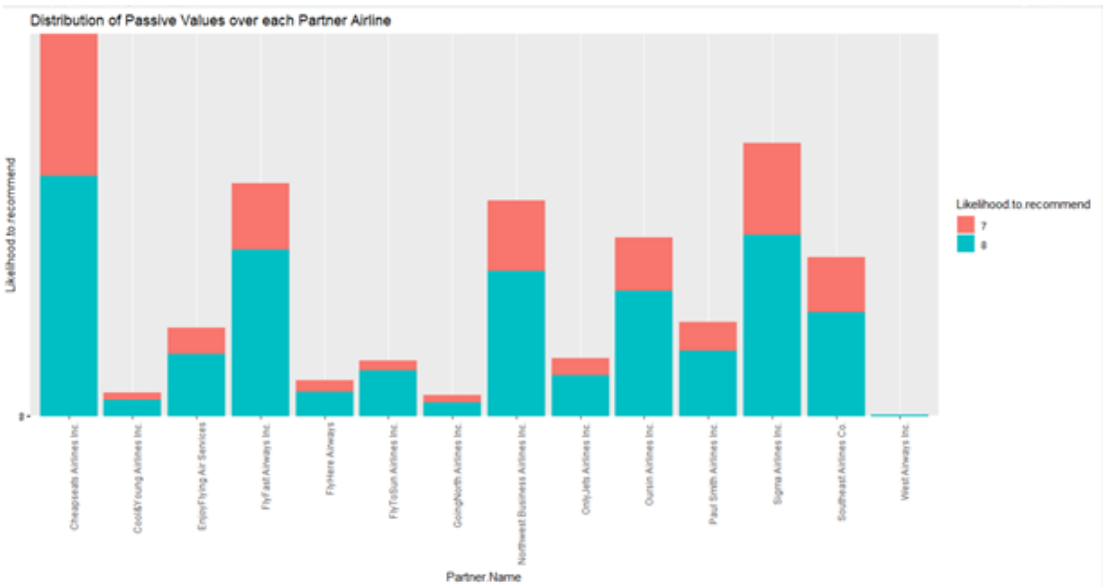

Figure 5 - Show the distribution of Likely.hood.to.recommend Passive values by airline

At this point it was determined that the most likely airline partners that have a significant sampling size of both Detractor and Promoters were: "Cheapseats Airlines Inc.", "FlyFast Airways Inc.", "Northwest Business Airlines Inc.", "Oursin Airlines Inc.", "Sigma Airlines Inc.", and "Southeast Airlines Co.". However, our group felt that it was important to create a sampling percentage analysis on the data in question to note what percentage of data would be neglected if the 8 non-selected airlines were left out of further modeling analysis.

The first step was to analyze the Detractor and Promotor dataframes to determine the appropriate airlines to use. Once that was determined we merged our dataframes (passive and detractor) and sub-set our data into the 14 data frames based on the 14 airline partners in order to find the total number of records (both passive and detractor) by airline. Finally, we divided this value by the total overall records of all Passive and Detractor records over all 14 airlines to come up with the percentage of relevant data recorded in each airline. As one can see from Table 1, 80% of the relevant data (passive and detractor records) reside in 6 of 14 airlines. Therefore, this proportion sampling analysis further justifies out previous distribution graph by showing that if one decides to only perform further analysis on "Cheapseats Airlines Inc.", "FlyFast Airways Inc.", "Northwest Business Airlines Inc.", "Oursin Airlines Inc.", "Sigma Airlines Inc.", and "Southeast Airlines Co." airlines an adequate representation of relevant data (80% of overall) will be present.

Combined Detractor/Passive Airline Partner Ratio		
Cheapseats Airlines Inc.	CAI_Ratio	0.22415
Cool&Young Airlines Inc.	CYAI_Ratio	0.01195
EnjoyFlying Air Services	EFAS_Ratio	0.04566
FlyFast Airways Inc.	FFAI_Ratio	0.13896
FlyHere Airways	FHA_Ratio	0.01823
FlyToSun Airlines Inc.	FTSAI_Ratio	0.02804
GoingNorth Airlines Inc.	GNAI_Ratio	0.01195
Northwest Business Airlines Inc.	NBAI_Ratio	0.11031
OnlyJets Airlines Inc.	OJAI_Ratio	0.03355
Oursin Airlines Inc.	OAI_Ratio	0.09851
Paul Smith Airlines Inc.	PSAI_Ratio	0.05056
Sigma Airlines Inc.	SAI_Ratio	0.14095
Southeast Airlines Co.	SAC_Ratio	0.08672
West Airways Inc.	WAI_Ratio	0.00046

Sum Taken	80%
Sum Dropped	20%

Table 1 shows the distribution of the combined Detractor and Passive records by airline. As one can see 80% of the meaningful data we deemed relevant for further analysis populates 6 out of the 14 airline partners.

At this point our group decided to only perform further analysis on the aforementioned 6 airlines therefore we created 6 dataframes for further analysis.

Key for the 6 datefrmes:

- dfCAI = "Cheapseats Airlines Inc."
- dfFFAI = "FlyFast Airways Inc."
- dfNBAI = "Northwest Business Airlines Inc."
- dfOAI = "Oursin Airlines Inc."
- dfSAI = "Sigma Airlines Inc."
- dfSAC = "Southeast Airlines Co."

Use of Modeling Techniques & Visualizations

LINEAR MODEL ANALYSIS

For the purposes of this linear model analysis I am going to convert the dataframe df column's Likelihood.to.recommend, Price.Sensitivity, and Year.of.First.Flight from type factor to num. This helps when plotting and analyzing the data in each analysis taken. The reset of the data that was analyzed remained the same type as discussed in the data munging section above.

Process for Linear Model Analysis

In order to analyze multiple linear model regression analysis a function was constructed in order to input a data frame which will return a linear model analysis based on the dependent

(likely.hood.to.recommend) based on the independent variables that were determined by our team as acceptable to analyze from the dataset provided. After the dependent variables were determined the group discussed the relevance of testing the entire airlines data frame given or subset the data somehow in order to break the analysis apart to provide a more focused understanding of the airline data. This process resulted in sub-setting the original data frame into 6 different data frames based on the variable airline partners (this process was discussed above in the Determined Appropriate Dataframes to Analyze section. Therefore the previously discussed linear modeling function was used to analyze all 6 subsets of the airline data. The summary of each of the 6 linear model function outputs were then returned as the functions output and an analysis of the functions Adjusted R-Squared value and Significant Variable (based on the coefficients calculated).

Result of the Linear Model Analysis

The linear model function for 12 of the 14 dataframes were determined using the same function (dependent/independent variables) designed. The dfCYAI data frame however did not have an adequate distribution of the Flight.Cancelled variable (i.e. 100% no cancelled flights) to provide a meaningful linear model output. Therefore this particular airline partner dataframes linear model analysis excluded the independent variable (cancelled flights) within its analysis. Additionally, the dfWAI dataframe only consisted of 6 records which resulted in inadequate overall amount of data to provide a meaningful linear model analysis. Therefore, our group excluded this airline partner within the final linear model analysis across all dataframes. Please note that because the airline partner only had limited recorders to analyze and after analyzing the (likely.hood.to.recommend) variable (50% Promoters, 33.33% Passive, 16.66% Detractors).

Based on the analysis for the 6 linear models it was determined that the significant variables were identified to be Airline.Status: Platinum, Airline.Status: Silver, Age, and Type.of.Travel: Personal Travel we determined these values to be the most significant over the others by creating a table which documented the results of all 6 linear model outputs. From this table the clear majority of overlapping significant variables where determined and documented above. Due to the length the entirety of the table was excluded from this report however Table 2 displays a shortened version of the table and the excel file of the full table will be included in the deliverables with this project for further analysis.

	Airline.StatusGold	Airline.StatusPlatinum	Airline.StatusSilver	Age	Price.Sensitivity	Type.of.TravelMileage tickets	Type.of.TravelPersonal Travel	ClassEco	Scheduled.Departure.Hour	Departure.Delay.in.Minutes	Arrival.Delay.in.Minutes	Flight.cancelled	Flight.Distance	Adjusted R-Squared
dfCAI		3	3	1	1		3							0.277
dfFFAI	3		3				3						1	0.3142
dfNBAI		1	3	2			3	1				3		0.269
dfOAI			3	1			3							0.2502
dfSAI		3	3	2		1	3		1					0.2581
dfSAC		2	3	2			3			1	3			0.2949
Key	3 (***)	Most Significant			dfCAI = Cheapseats Airlines Inc.		dfOAI = Oursin Airlines Inc.							
	2 (**)	Significant			dfFFAI = FlyFast Airways Inc.		dfSAI = Sigma Airlines Inc.							
	1 (*)	Less Significant			dfNBAI = Northwest Business Airlines Inc.		dfSAC = Southeast Airlines Co.							

Table 2 - Shortened output table analysis of the linear model's taken place from the R coded output

The Adjusted R-Squared values for our linear model analysis were 0.277 for "Cheapseats Airlines Inc.", 0.3142 for "FlyFast Airways Inc.", 0.269 for "Northwest Business Airlines Inc.", 0.2502 for "Oursin Airlines Inc.", 0.2581 for "Sigma Airlines Inc.", and 0.2949 for "Southeast Airlines Co." which noted from "An Introduction to Data Science" by J. Saltz and J. Stanton (page 203) when analyzing human behavior (which we are in this project) a very good R-Squared values can be between (0.2 and 0.3) due to the fact that humans are notoriously unpredictable. Therefore it is not alarming to our group to have Adjusted R-Squared values between 0.25-0.31 as shown above.

Now that we have looked at the 6 linear models and have determined that the significant variables are Airline.StatusPlatinum, Airline.StatusSilver, Age, Type.of.TravelPersonal Travel the only remaining thing to determine through linear modeling is which age is most beneficial to raise our detractor value, therefore, using ggplot we can graph the dependent variable (Likelihood.to.recommend) vs. Age and using the stat_smooth function to create a linear model tread line to plot a linear model line to the graph.

By doing this we see that the younger one is the more likely a higher passive or detractor score was present, therefore, we recommend focusing marketing campaigns to appeal to Younger (below 45), Platinum or Silver airline status members, who are looking to travel on personal travel in order to increase Detractor and Passive Likelihood.to.recommend scores in order to increase the overall NPS.

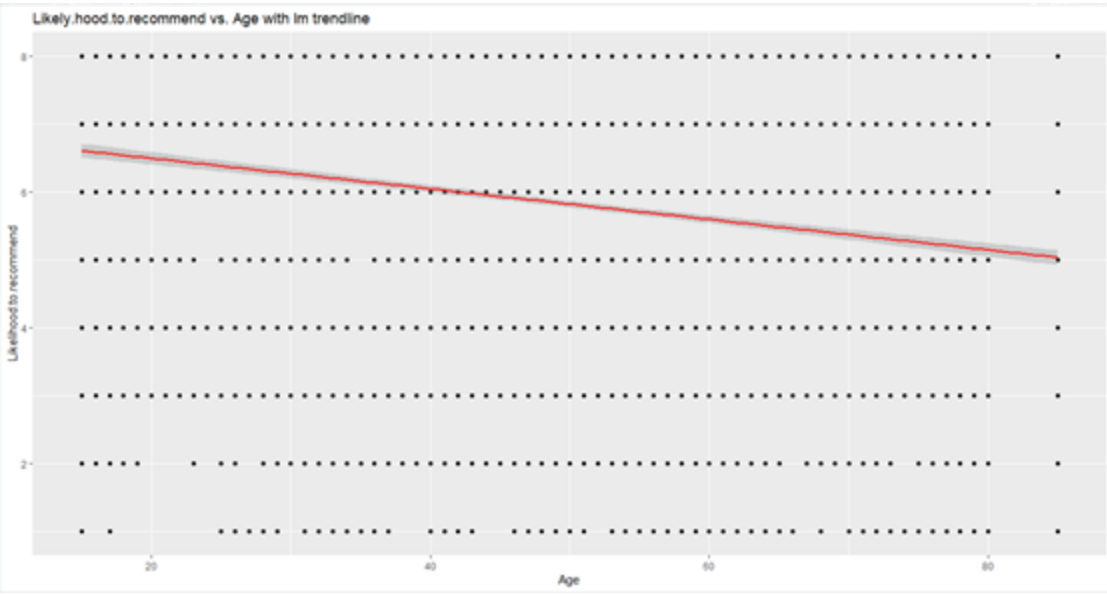

Figure 6 - Fitting a linear model trend line for Likelihood.to.recommend vs. Age over all 6 airlines

Sanity Checks / Verification

In an attempt to check over the above linear analysis we choose to combine the dataframes of all 6 airlines into one overall data frame (PD_Airlines) and run a linear model analysis on its entirety. The models output can be viewed in Figure 7. As one can see all the above significant variables identified as the airline companies recommended focus areas in the linear model section match the analysis of each separate linear model identified (Airline.Status Platinum, Airline.StatusSilver, Age, Type.of.TravelPersonal). However, this analysis does include a few additional variables that were noted in the previously discussed Excel table calculations for identifying the significant variables to determine the focus areas for the company's executives. These variables were identified in the previous analysis as mentioned but were purposely excluded because they were determined to be less significant across all 6 airlines then the above identified. These variables are as followed: Significant Variable (***): Airline.StatusGold; less significant variable (**): Scheduled.Departure.Hour, Flight.cancelledYes, and minor significance variables: (*) Type.of.TravelMileage tickets, and Arrival.Delay.in.Minutes. The calculated significance table can be identified on the excel file labeled LinearModelAnalysis.xlsx. Finally, as one can see from the Figure 7 model the Adjusted R-Squared value was 0.2681 which is acceptable when modeling human behavior.

```
> LinModel(PD_Airlines)

Call:
lm(formula = Likelihood.to.recommend ~ Airline.Status + Age +
    Gender + Price.Sensitivity + Year.of.First.Flight + Flights.Per.Year +
    Loyalty + Type.of.Travel + Total.Freq.Flyer.Accts + Shopping.Amount.at.Airport +
    Eating.and.Drinking.at.Airport + Class + Day.of.Month + Scheduled.Departure.Hour +
    Departure.Delay.in.Minutes + Arrival.Delay.in.Minutes + Flight.cancelled +
    Flight.time.in.minutes + Flight.Distance, data = Inputdf)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7548 -0.9552  0.2178  1.3025  3.9960

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   25.5495124 15.6184692   1.636 0.101931
Airline.StatusGold             0.3574209  0.1001536   3.569 0.000362 ***
Airline.StatusPlatinum        -0.9058447  0.1430026  -6.334 2.58e-10 ***
Airline.StatusSilver           1.3401173  0.0683100  19.618  < 2e-16 ***
Age                           -0.0088390  0.0015071  -5.865 4.77e-09 ***
GenderMale                    -0.0232570  0.0487545  -0.477 0.633366
Price.Sensitivity             -0.0680253  0.0416205  -1.634 0.102232
Year.of.First.Flight          -0.0091783  0.0077840  -1.179 0.238405
Flights.Per.Year              -0.0032278  0.0022331  -1.445 0.148394
Loyalty                       -0.0467367  0.0651502  -0.717 0.473179
Type.of.TravelMileage tickets  0.2082421  0.0941132   2.213 0.026963 *
Type.of.TravelPersonal Travel -1.4938496  0.0537364 -27.800  < 2e-16 ***
Total.Freq.Flyer.Accts        -0.0421514  0.0242386  -1.739 0.082091 .
Shopping.Amount.at.Airport     0.0008688  0.0004586   1.895 0.058204 .
Eating.and.Drinking.at.Airport 0.0007744  0.0004097   1.890 0.058807 .
ClassEco                      -0.0231903  0.0911126  -0.255 0.799101
ClassEco Plus                  0.0780681  0.1134240   0.688 0.491303
Day.of.Month                   0.0030289  0.0026582   1.139 0.254575
Scheduled.Departure.Hour      -0.0138005  0.0049955  -2.763 0.005754 **
Departure.Delay.in.Minutes    -0.0009935  0.0020404  -0.487 0.626336
Arrival.Delay.in.Minutes      -0.0045732  0.0020215  -2.262 0.023724 *
Flight.cancelledYes            0.4770426  0.1456808   3.275 0.001065 **
Flight.time.in.minutes        -0.0008314  0.0012354  -0.673 0.500969
Flight.Distance                0.0001905  0.0001478   1.289 0.197416
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.663 on 5195 degrees of freedom
Multiple R-squared:  0.2713,    Adjusted R-squared:  0.2681
F-statistic: 84.11 on 23 and 5195 DF,  p-value: < 2.2e-16
```

Figure 7 - Output of the sanity check model for the entire filtered airline (combination of 6 dataframes)

Finally, it has been noticed that there is a connection with the final results taken from the above two linear analysis with that of other analysis (i.e. associated rule mining analysis) determined in this project. This means that the four identified important independent variables determined in our linear analysis (Airline.Status Platinum, Airline.StatusSilver, Age, Type.of.TravelPersonal) also reflect similar dependencies on likely.hood.to.recommend from other analysis using

different types of model. The overall identification of the similarity adds confidence to our overall analysis of this project.

ASSOCIATION RULES

To use the Apriori() function on the data, we had to first convert variables of Type.of.Travel, classes Flight.cancelled, Status, Likelihood.to.recommend, gender, age and class to factors so that they could be put into a dataframe together that could then be converted to a matrix. We chose only variables that made the most sense to assess with Apriori() function because it is unfeasible to try to convert all columns in the dataframe to one single mode, which is required to create a matrix. In retrospect, it could have worked to bin()/group all of the variables in the dataset to convert them all to factors to run the Apriori() function. However, choosing a smaller subset based on the results of the linear regression as well as intuition so that we could get as granular as possible with the results.. The association rules portion focuses on the passive and detractor traits to complement the outcomes from the linear regression and generate rules that describe which variables occur together regularly.

Process for Association Rules

We extracted columns of interest from the cleaned dataframe to create matrices of transactional data for each of the airlines that were determined to contain more than 80% of valuable responses while throwing out airlines weh insufficient responses and small sample size or unusable results. For the association rules we converted the likelihood.to.recommend column to a range of detractor(0-6), passive(7-8), and promoter(9-10) using ifelse() command so that running the Apriori() function generates rules that describe variable impact on the NPS. We made the same conversion for age so that it could be read as a factor and used in the new matrix. The age levels are senior (50+ years old), Adult (18-50 years old), and youth (1-17years old).

We used piping to pass the data from the cleaned dataframe into a new dataframe, selected only the factor variables that are listed above, and filtered by Partner.name to get the data for each individual airline. Once the new, airline-specific dataframes were made, the last step was to convert them into transactional matrices. After all columns were converted, we converted the dataframes for each airline into matrices and ran the Apriori() function to generate rules.

For running the function, we selected a relatively low number for support (around .005-.1) because we do not necessarily care how uncommon a variable is if it has a high lift or always has an impact on our dependent variable. If a pairing only occurs a few times with the Likelihood.to.reccomend, then we do want to ignore it, but a support of .005-.1 usually allowed a good selection of reasonably common variables.I chose a higher number for confidence (.3-.8 depending on variable frequency within the airline). We generally adjusted confidence and support to strike a balance between being inclusive and *not* generating too many rules to assess.

The last part of setting up the Apriori() function is choosing the left hand side (LHS) and the right hand side (RHS). We chose the default for the LHS so that the function would compare all the selected variables with Likelihood.to.recommend, which we defined as the RHS because it is our dependent variable.
We ran the function twice for each airline. The first time to find the variables that occur with the detractor scores and then again for the variables that occur with passive scores. In the end, many of the results were similar enough that a clean matrix of all the data for all six airlines would give the same level of information while allowing a more granular assessment of individual variables.

Apriori Results

For the next step the above Apriori() strategy was applied again, but this time filtered by all six airline partners so that they could all be accessed in the same matrix. Before running the ApriorиI() function on the new matrix, we created a histogram to view the variable frequencies and determine that we should use a support value of .05 to include even the least common variables. The bar chart is shown below.

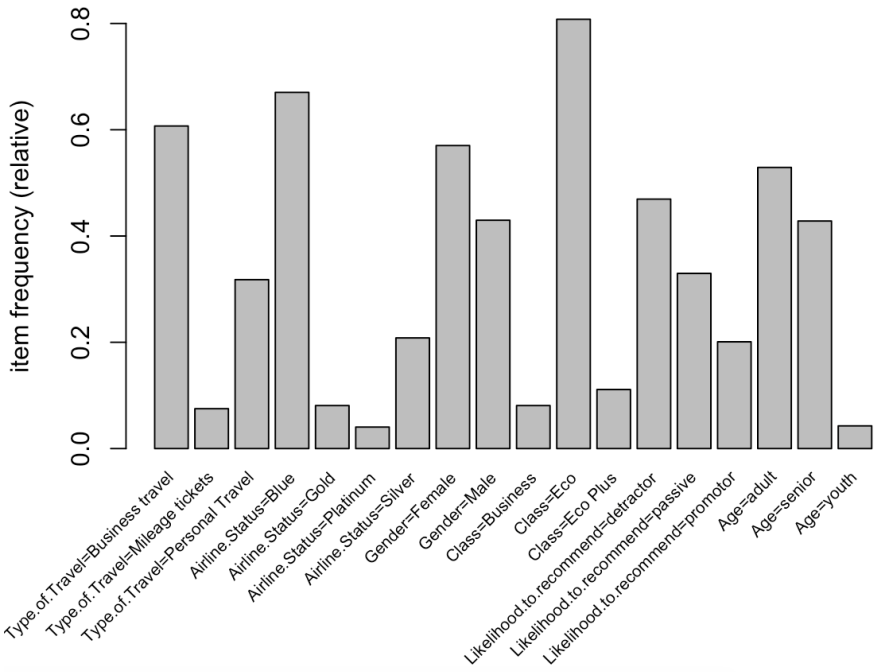

Figure 8 - Variable frequency from the matrix of six airlines' important data

After setting the support, we chose a confidence level of .3, which returned 70 rules. To better understand the rules, we created a scatterplot to see where they fell on a scale of lift as shown below.

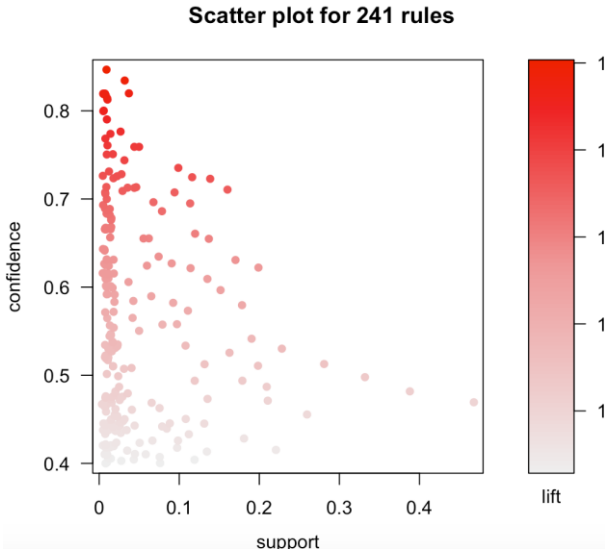

Figure 9 - Scatter plot of confidence, support, and lift for variables in the new matrix

We can see from the scatterplot that there are a significant number of rules falling above 1.2, so we adjusted our Apriori() function to only include rules with a lift above 1.2 to get a more manageable set of only 24 rules. We repeated the process for the passive group to get a set of 21 rules. Inspecting the rulesets showed that the most interesting variables over the six important airlines were still Type.of.Travel, Airline.Status, and Age. The detractors traits with the best ratios of support to confidence to lift were Age=Senior and Type.of.Travel=personal. Passive customers were most likely to have Airline.Status=Silver and Age=Adult.

To go a step farther, we ran the apriori() function again to consider both detractors and passives, but to focus on just of of the three interesting variables. The function was applied to one variable at a time. For age, seniors have the most lift and confidence meaning that they are the most likely to give a detractor review. Passive customers are most likely to be Adults by a small margin, but many customers in the adult category also gave passive scores, so they are not as important a group to focus on.

For airline status, detractors are most likely to have Platinum Status. However, the detractors with Platinum Status have a very low support, so if we consider only variables with support above .05, the Blue Status passengers become the most common Detractors. Passive customers are most likely to have Silver Status. Blue Status is by far the most prolific across all customers, so it is important to note that this variable had a sufficiently high lift and confidence to suggest that it is not a coincidence due to the variable's high frequency. While it provides a good, large group of customers to focus on, some of the relationship occurs as a result of the scarcity of customers with Gold or Platinum Status. However, it may still be worth looking into why Gold and Platinum status holders are often detractors. For type of travel, detractors usually have a personal travel type.

In conclusion, the association rules suggest that Southeast Airlines Inc. should focus on customers who have a personal travel type, are seniors, and have a blue status if they wish to focus on a detractor profile. For focusing on passive travelers, the airline should look to adult customers with silver status.

DECISION TREE AND RANDOM FOREST ANALYSIS

Process and Results for Decision Tree and Random Forest Analysis

Decision Tree Analysis

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. At this case, decision tree would be a good method to find the important variables of improving the net promoter score. Also, the decision tree model could validate the results of other models used in this case like linear model.

To use the decision tree model, we divided the *likelihood.to.recommend* column into 3 different factors, 'Detractor', 'Passive', 'Promoter'.

```
 for(i in seq(1,length(Inputdf[,1]))){
  if(Inputdf$Likelihood.to.recommend[i] <= 6){
    Inputdf$Likelihood.to.recommend[i] <- 'Detractor'
  }else if(Inputdf$Likelihood.to.recommend[i] >=9){
    Inputdf$Likelihood.to.recommend[i] <- 'Promoter'
  }else{
    Inputdf$Likelihood.to.recommend[i] <- 'Passive'
  }
Inputdf$Likelihood.to.recommend <- as.factor(Inputdf$Likelihood.to.recommend)
```

Then we removed all the geographical columns like *Destination.City, Origin.City* since they would influence the result of the decision tree analysis. After the processing of data frame, we used the caret package to create training and test sets.

```
#install.packages('caret')
library(caret)
#use caret package to create training and test sets

#makes the sampling predictable
set.seed(111)
str(prac_df)
# Randomly sample elements to go into a training data set
trainList <- createDataPartition(y=prac_df$Likelihood.to.recommend,p=.70,list=FALSE)
```

```
# Include all of those elements in the training set
trainSet <- prac_df[trainList,]

# Construct test set from everything that didn't go into the training
testSet <- prac_df[-trainList,]
```

We used the *rpart* package in R to build the decision tree model.

```
# Use the trainSet
like_tree <- rpart(Likelihood.to.recommend~.,data = trainSet)
# Set the minimum number of observations that must exist in a node in order for a split to be attempted
control = rpart.control(minsplit = 8)
like_tree
```

After building the model, we used the *fancyRpartPlot* function from *rattle* package to create a fancy visualization of this decision tree model.

```
#Pretty visualization using fancyRpartPlot function
library(rattle)
library(RColorBrewer)
fancyRpartPlot(like_tree)
```

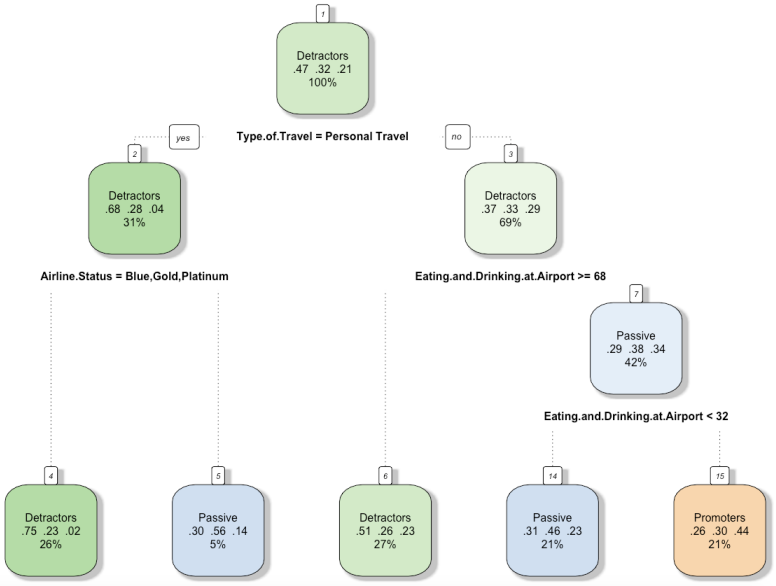

Figure 10 - Visualization of decision tree model of whole dataset

The decision tree model is a model with binary response. Left nodes show all 'yes' results of condition, and right nodes show all 'no' results of condition. Each node shows three things: the predicted class (Detractor, Passive or Promoter), the predicted probability of these classes, the percentage of observations in the node.

In this model, we could see the main logical conditions are mainly in *Type.Of.Travel, Eating.and.Drinking.at.Airport, Airline.Status* variables.

Then we used the varImp function to get the variable importance of each attribute, and got a bar chart of variable importance.

```
vi <-varImp(like_tree)
vi$variables <- rownames(vi)
# using ggplot2 draw a bar chart of variable importance
vi<- vi[vi[,1]>0,]
# using RColorBrewer to set the color
mycolor <- brewer.pal(8,'GnBu')
plt_variance <- ggplot(data =vi, aes(reorder(x=vi[,2],vi[,1]),y=vi[,1])) + geom_bar(stat =
'identity',fill = mycolor) + theme(axis.text.x = element_text(angle = 90, hjust = 1))
plt_variance
```

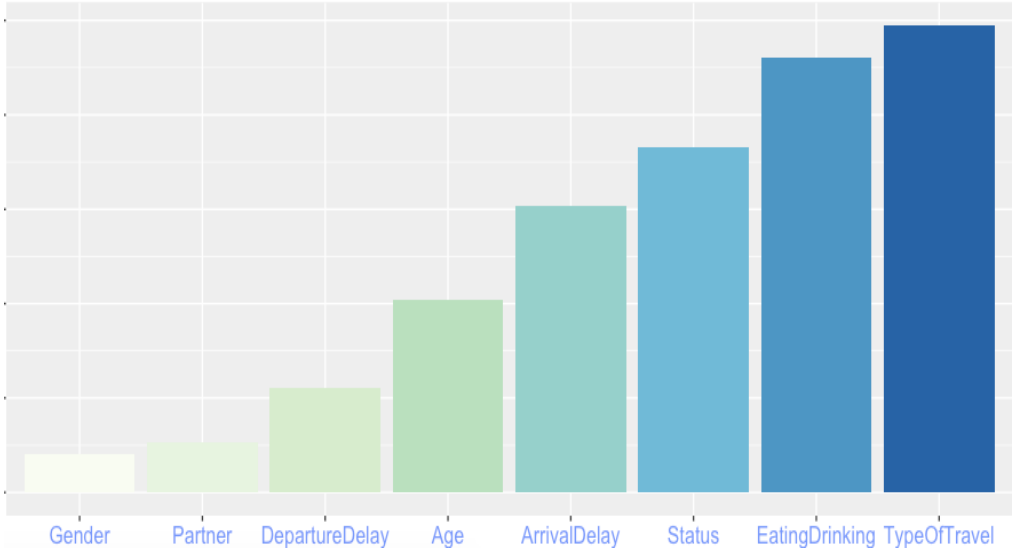

Figure 11 - Variable importance plot of the model

An overall measure of variable importance is the sum of the goodness of split measures for each split for which it was the primary variable. And the variable importance shows which attribute is more important to the *likelihood.to.recommend* score.
From the above figures we could see the *TypeOfTravel, Eating and Drinking at Airport, Airline.Status* have the highest variable importance.

Then we used the model to predict the testSet and created a confusion matrix.

```
# Use the testSet to evaluate the like_tree
predictValues <- predict(like_tree,newdata=testSet,type = "class")
```

```
#simpler to do confusion matrix
#install.packages('e1071') in order to create the confusion matrix
library(e1071)
confusion <- confusionMatrix(predictValues,testSet$Likelihood.to.recommend)
confusion
```

	Class: Detractors	Class: Passive	Class: Promoters
Sensitivity	0.7201	0.3275	0.42792
Specificity	0.6051	0.7996	0.84695
Pos Pred Value	0.6173	0.4313	0.43185
Neg Pred Value	0.7097	0.7194	0.84486
Prevalence	0.4693	0.3169	0.21375
Detection Rate	0.3380	0.1038	0.09147
Detection Prevalence	0.5475	0.2407	0.21181
Balanced Accuracy	0.6626	0.5636	0.63743

Figure 12 - Output decision tree model confusion matrix

The prediction accuracy for these three classes are 0.6626, 0.5636, 0.63743.

Random Forest Analysis

A Random Forest is essentially a collection of Decision Trees. A decision tree is built on an entire dataset, using all the features/variables of interest, whereas a random forest randomly selects rows and specific variables to build multiple decision trees from and then averages the results. The random forest algorithm has better performance than single decision tree in a general way.

Here we used the *RandomForest* package in R to predict the *likelihood.to.recommend* column.

```
#install.packages('randomForest')
library(randomForest)
library(dplyr)
#convert all the character columns to factor
trainSet=trainSet %>% mutate_if(is.character, as.factor)
testSet=testSet %>% mutate_if(is.character, as.factor)

#build a random forest model of 300 trees
like_rm_tree <- randomForest(Likelihood.to.recommend~.,data = trainSet,ntree=300)

#plot the importance
varImpPlot(like_rm_tree)
```

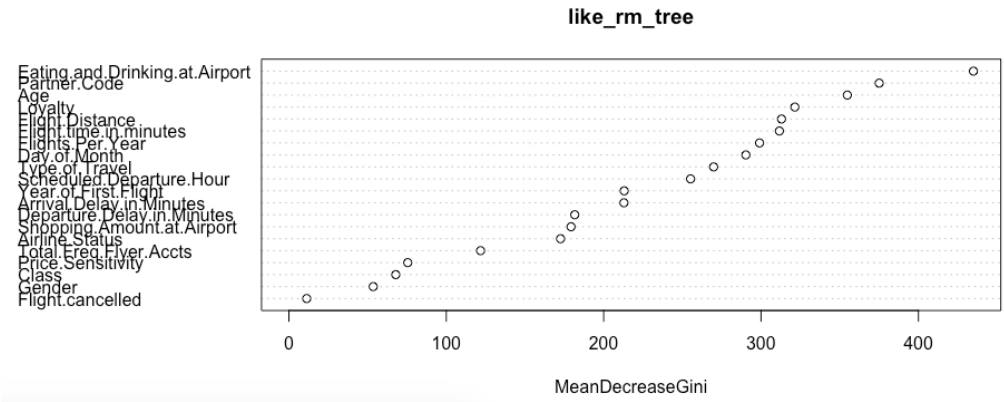

Figure 13 - MeanDecreaseGini plot of the random forest model

From the random forest variable importance (MeanDecreaseGini) plot we could see that the *Eating.and.Drinking.at.Airport, Partner.Code, Age* are the main factors of influencing the nps score.

Also, we get the confusion matrix of the random forest model.

```
#predict the testSet
rf_pre <- predict(like_rm_tree,newdata=testSet,type = "class")
#simpler to do confusion matrix
confusion <- confusionMatrix(rf_pre,testSet$Likelihood.to.recommend)
confusion
```

```
Statistics by Class:

                     Class: Detractors Class: Passive Class: Promoters
Sensitivity                     0.7574         0.4074          0.35205
Specificity                     0.5740         0.7944          0.90635
Pos Pred Value                  0.6113         0.4789          0.50545
Neg Pred Value                  0.7279         0.7429          0.83727
Prevalence                      0.4693         0.3169          0.21375
Detection Rate                  0.3555         0.1291          0.07525
Detection Prevalence            0.5816         0.2695          0.14888
Balanced Accuracy               0.6657         0.6009          0.62920
```

Figure 14 - Output random forest model confusion matrix

The balanced accuracy is higher than the decision tree. However, in this case, predicting the user nps score is not the primary task. Instead, the main task is to find the main factors relevant to the nps score and find the insights behind them. For further analysis, we still used the decision tree model and we could decrease the computing time without missing any important information.

Further Decision Tree Analysis

Based on the findings of decision tree and random forest analysis above. We decided to do further analysis for different airlines, different age groups and different types of travel.

To make the further analysis more convenient, we wrote a function to build a decision tree model and output the visualization of the model and the variable importance plot. And then we just need to input different dataset to get all the information we want.
(processing part is in the code file)

```
DecisionTreeModel <- function(Inputdf){
 # Get the train set and test set
  trainList <- createDataPartition(y=Inputdf$Likelihood.to.recommend,p=.70,list=FALSE)
  trainSet <- Inputdf[trainList,]
  testSet <- Inputdf[-trainList,]
 # Build the model
  like_tree <- rpart(Likelihood.to.recommend~.,data = trainSet)
  control = rpart.control(minsplit = 8)
 # Get the visualization of the model
  prp(like_tree, faclen = 0, cex = 0.8, extra = 1)
  predictValues <- predict(like_tree,newdata=testSet,type = "class")
 # Get the confusion matrix
  confusion <- confusionMatrix(predictValues,testSet$Likelihood.to.recommend)
  print(confusion)
  # Explore the variable importance
  vi <-varImp(like_tree)
  vi$variables <- rownames(vi)
  vi<- vi[vi[,1]>0,]
  plt_variance <- ggplot(data =vi, aes(reorder(x=vi[,2],-vi[,1]),y=vi[,1])) + geom_bar(stat =
'identity') + theme(axis.text.x = element_text(angle = 90, hjust = 1))
  plt_variance
}
```

Analysis of Different Airlines

The linear model chose the most valuable 6 airlines. (*"Cheapseats Airlines Inc.", "FlyFast Airways Inc.","Northwest Business Airlines Inc.", "Oursin Airlines Inc.","Sigma Airlines Inc.","Southeast Airlines Co."*)
To gain the insights of each airline, we run the decision tree model for each airline.

```
#Do decision tree analysis for each airline and observe the results
#"Cheapseats Airlines Inc."
DecisionTreeModel(dfCAI)
# "FlyFast Airways Inc."
```

```
DecisionTreeModel(dfFFAI)
#"Northwest Business Airlines Inc."
DecisionTreeModel(dfNBAI)
#"Oursin Airlines Inc."
DecisionTreeModel(dfOAI)
#"Sigma Airlines Inc."
DecisionTreeModel(dfSAI)
#"Southeast Airlines Co."
DecisionTreeModel(dfSAC)
```

And then we collected the top 3 important factors of each airline

CAI	Type.Of.Travel	Airline.Status	Eating.and.Drinking
FFAI	Type.Of.Travel	Shopping Amount	Eating.and.Drinking
NBAI	Eating.and.Drinking	Type.Of.Travel	Airline.Status
OAI	Eating.and.Drinking	Departure.Delay	Airline.Status
SAI	Eating.and.Drinking	Type.Of.Travel	Arrival.Delay
SAC	Eating.and.Drinking	Age	Airline.Status

Besides the *TypeOfTravel, Eating and Drinking at Airport, Airline.Status* factors, the *Shopping Amount, Departure Delay, Arrival Delay, Age* are also important variables depend on different airlines.

For example, for FFAI(***FlyFast Airways)*** airline, improving the shopping service at the airport may improve the nps score efficiently. And for OAI(***Oursin Airlines)***, SAI(***Sigma Airlines)***, improving the management of flight schedule and reduce the delay time may help to improve the nps score.

Analysis of Different Age Groups

Based on the finding of decision tree and random forest analysis above, the Age attribute is an important factor that could not be ignored. And we decided to run models on different age groups.

```
#Still use the function we created
df_senior <- df[df$Age >= 65,]
df_young <- df[df$Age < 65,]
DecisionTreeModel(df_senior)
DecisionTreeModel(df_young)
```

- **For age group(>65):**

The visualization of the decision tree and variable importance plot shown below

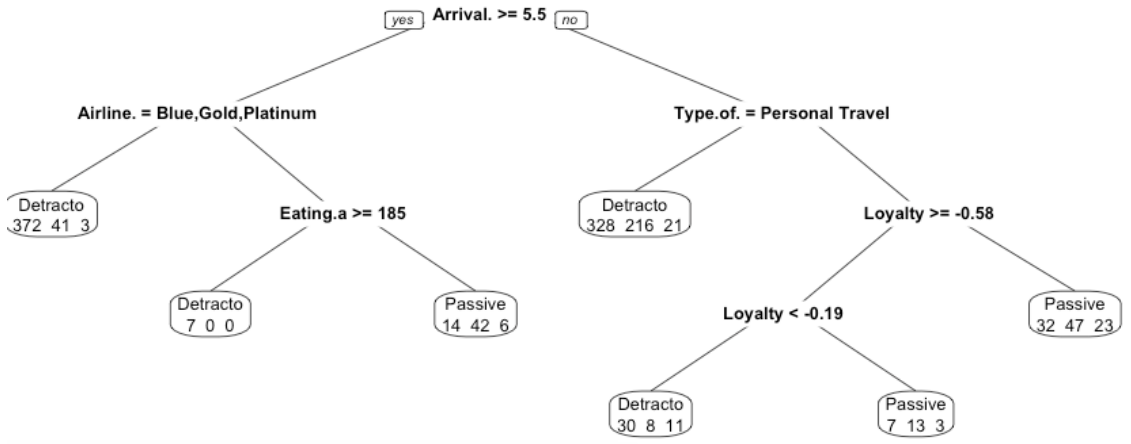

Figure 15 - Visualization of decision tree model for age group(>65)

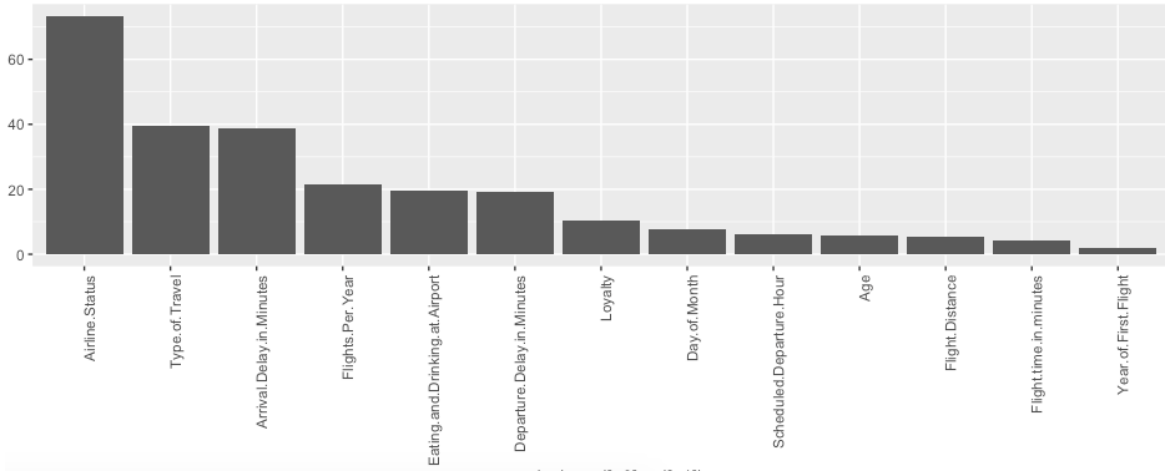

Figure 16 - Variable importance plot of the model for age group(>65)

We could see for the (age>65) group, the *Arrival.Delay.in.Minutes* is a significant variable. From the visualization, we could see that if the arrival delay is more than 5.5 minutes, then almost 90% percent of people would become detractor which means their *likelihood.to.recommend* score would be below 7.

- **For age group(<65):**

The visualization of the decision tree and variable importance plot shown below.

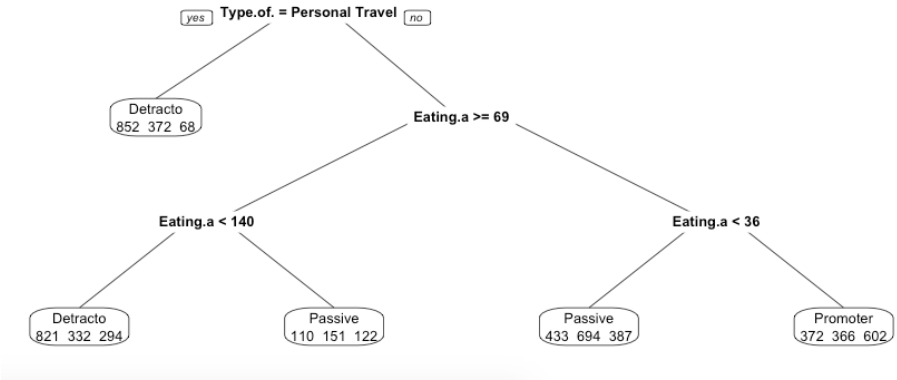

Figure 17 - Visualization of decision tree model for age group(<65)

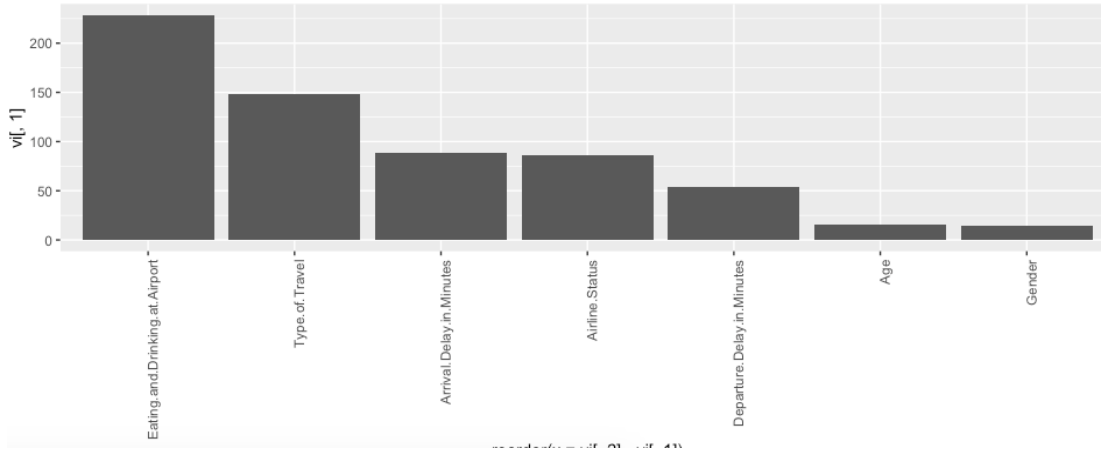

Figure 18 - Variable importance plot of the model for age group(<65)

We could see that the *Eating.and.Drinking.in.Airport* is so important for this age group that that factor is shown in most nodes of the tree and has a huge variable importance. To improve the nps score of this age group, we should pay more attention to improve the eating and drinking experience in the airport.(especially large amount of consumption).

Analysis of Different Types of Travel

Based on the finding of decision tree and random forest, even in the analysis of different airlines and age groups, the *Type.Of.Travel* influences the nps score very much.
So we also do the analysis for different types of travel.

```
df_bus <- df[df$Type.of.Travel == 'Business travel',]
df_per <- df[df$Type.of.Travel == 'Personal Travel',]
df_mil <- df[df$Type.of.Travel == 'Mileage tickets',]
DecisionTreeModel(df_bus)
DecisionTreeModel(df_per)
DecisionTreeModel(df_mil)
```

- **Business Travel**

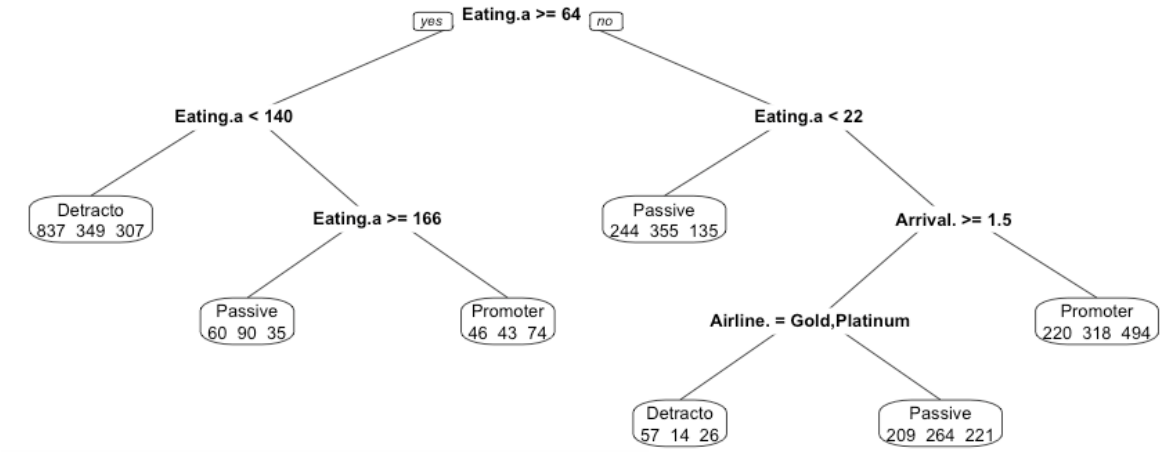

Figure 19 - Visualization of decision tree model for business travel type of travel

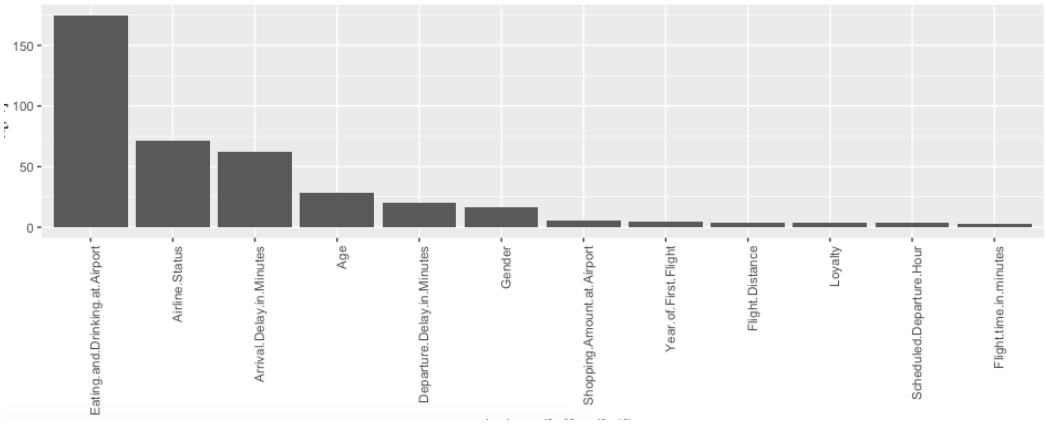

Figure 20 - Variable importance plot of the model for business travel type of travel

We could see Eating.and.Drinking in airport is still the main factor. Most customers who spent more than 64 dollars eating and drinking in airport gave a low likelihood.to.recommend score. Based on that, we need to focus on the service quality of eating and drinking at the airport if we want to improve the nps score of business people.

- **Mileage Tickets**

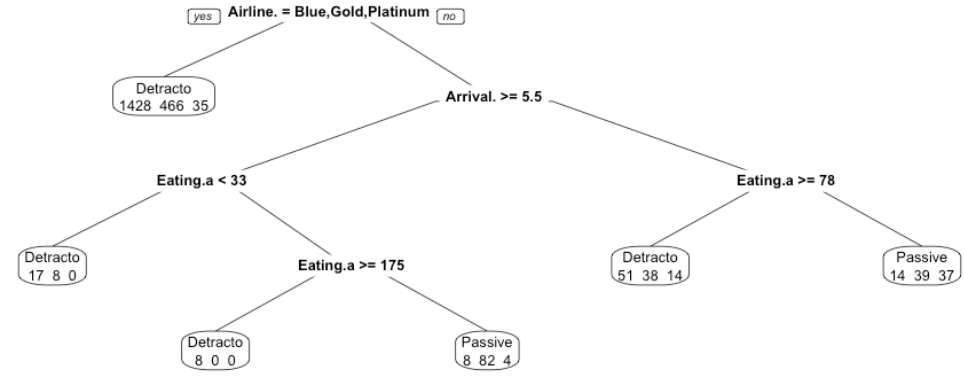

Figure 21 - Visualization of decision tree model for mileage tickets type of travel

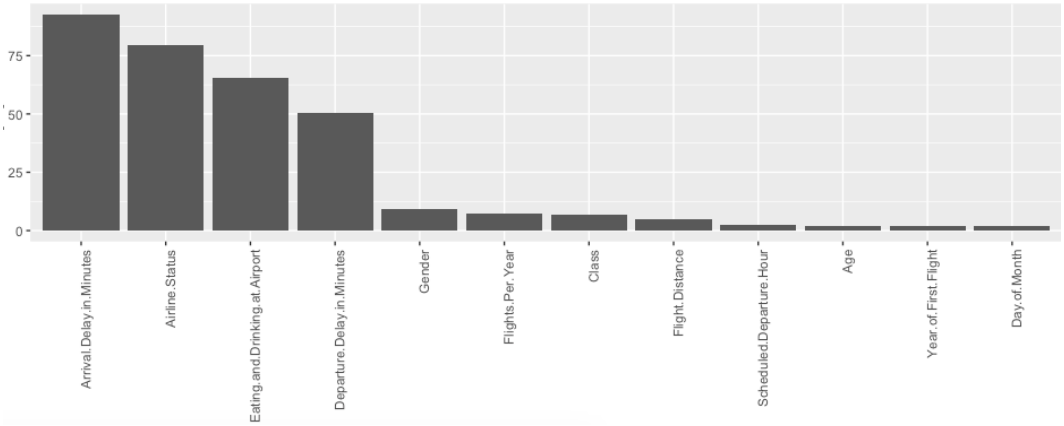

Figure 22 - Variable importance plot of the model for mileage tickets type of travel

For mileage tickets passengers, still only *Silver* status gave a high score of recommend. Another important factor is the arrival delay time.

- **Personal Travel**

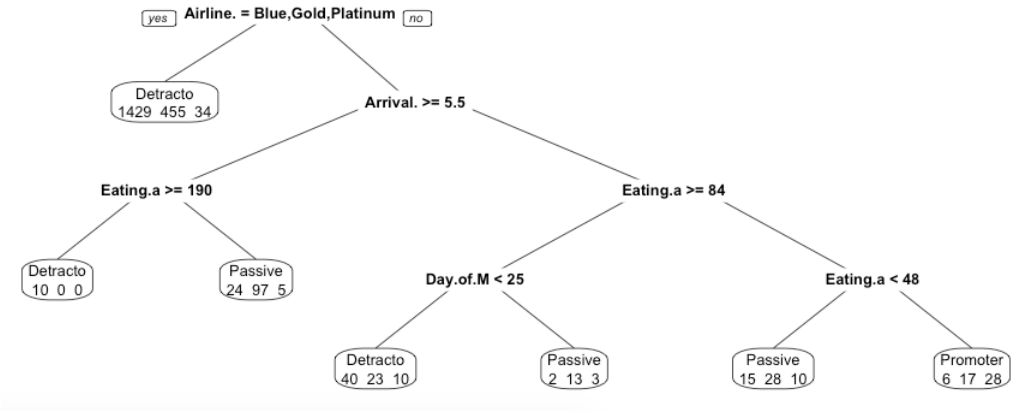

Figure 23 - Visualization of decision tree model for personal travel type of travel

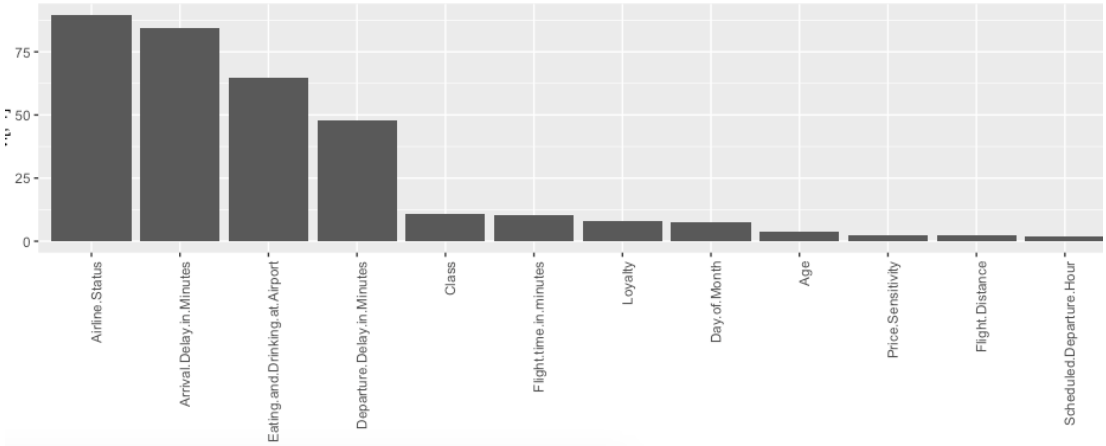

Figure 24 - Variable importance plot of the model for personal travel type of travel

For personal travel passengers, only *Silver* status gave a high score of recommend. We need to do more survey of *Blue,Gold,Platinum* status passenages to figure out why they would give a low recommend score.

SENTIMENT ANALYSIS

Process for Sentiment Analysis

We conducted a sentiment analysis of the customer submitted comments from within the Free Text column of the data. This was first done on the complete Free Text data available for an overall analysis. In order to conduct this analysis we created a value containing the Free Text data. We then created positive and negative word dictionaries using positive and negative word text files. We then created a words vector source as well as a words corpus suing the Free Text data, followed by the creation of a term document matrix. From here we were able to create a word cloud using frequencies of words within the data. Additional code was then added to create a more appealing looking word cloud. From here we then proceeded to the actual sentiment analysis. This was performed by summing the amount of words in the data, matching those words to the words in the positive and negative dictionaries respectively, and calculating the total number of positive and negative words in the data. The number of positive or negative words was then divided by the total number of words to calculate the percentage of positive words and the percentage of negative words in the data. After calculating these things for the overall Free Text data, we then replicated the steps above but for each of the six individual airline that we were focusing on for our analysis. This was done by separating the complete data frame into smaller tribbles based on the different airline partners, then by using those tibbles to create a new data frame for each airline partner. After this was done the steps described above were performed on each of the individual airline partners' Free Text data.

Result of the Sentiment Analysis

Our analysis resulted in the percentage of positive words and the percentage of negative words both overall and by airline partner, as well as word cloud generation for the overall data and the data by airline partner. The overall sentiment analysis resulted in more positive words at 31 percent positive. In the analysis per airline we discovered only two out of the six airlines focused on had a more negative percentage. The other four airlines had a more positive percentage of words. These results are listed below:

Overall Sentiment: 31 percent positive words, 24 percent negative words

Negative Leaning Sentiments

- Cheapseats Airlines Inc.
    - 12 percent positive words

    - **13 percent negative words**
- FlyFast Airways Inc.
    - 10 percent positive words
    - **14 percent negative words**

Positive Leaning Sentiments

- Northwest Business Airlines Inc.
    - **13 percent positive words**
    - 9 percent negative words
- Oursin Airlines Inc.
    - **22 percent positive words**
    - 11 percent negative words
- Sigma Airlines Inc.
    - **16 percent positive words**
    - 10 percent negative words
- Southeast Airlines Co.
    - **19 percent positive words**
    - 12 percent negative words

Based on these results we generated word clouds for the two airline partners with negative leaning sentiment to gain further insights into what those customers were most focused on, and what gave us the most potential for actionable recommendations. These word clouds can be seen below:

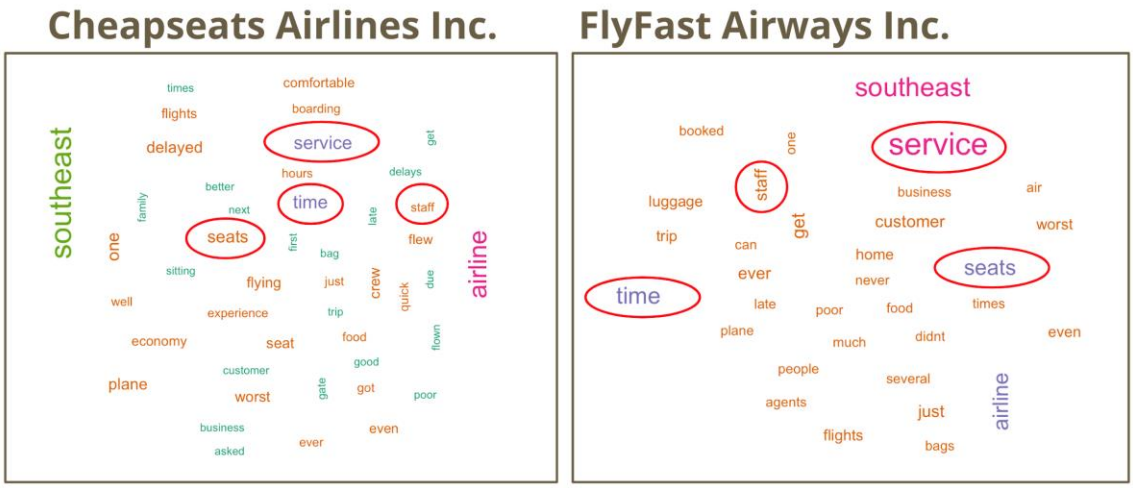

Figure 25 - Word clouds generated for free text data within Cheapseats Airlines Inc. and FlyFast Airways Inc.

From these word clouds we were able to identify the four most frequently mentioned terms across both airlines that could provide actionable insights. These were service, time, seats, and staff.

Actionable Insights / Overall Interpretation of Results

Justification

The strategy was to choose modeling techniques that would complement each other. The linear regression told us which variables were statistically significant and the association rules modeling supported that result by showing that the same important variables occur with Likelihood.to.reccomend. The linear modeling told us that Airline.StatusPlatinum, Airline.StatusSilver, Age, and Type.of.TravelPersonal Travel were important variables. The association rules agreed with that outcome and gave the added insight that the Platinum Status was a very small group of customers, making them slightly less important.

The decision tree acted to capture a different kind of information that the linear modelling and association rules does not show. It added another layer by considering all of the variables and all of the levels of Likelihood.to.recommend to give added insight into consumer behavior.

Finally, sentiment analysis addressed the sparse, but important comments left by people. It is worthwhile to find patterns in the comments because comments are usually on topics that are of the most significance to customers and it adds a helpful human element to an otherwise dry statistical extrapolation of a customer profile. All of the models complement each other to generate a more reliable and complete analysis.

Conclusion

The linear model analysis that was determined from the sub-set of meaningful munged data of passive and detractor records provided to group 2 from the Southeast Airline Company found that the independent variables (Airline.Status Platinum, Airline.StatusSilver, Age, Type.of.TravelPersonal) reflected significance to our dependent variable likely.hood.to.recommend. This was important to understand in order to best recommend market analysis for Southeast Airlines in order to determine the best methods to understand passive and detractor influence. In turn once these influences are understood Southeast Airlines can work to eliminate said influence to achieve a rise in overall passives and detractors values which in turn will increase the overall NPS score for the company.

The association rules re-affirm and deepen the explanation from the linear modeling. They suggest that Southeast Airlines Inc. should focus on customers who have a personal travel type, are seniors, and have a blue airline status if they wish to focus on a detractor profile. For focusing on passive travelers, the airline should look to adult customers with silver status. Finding out what will sway these passengers to become either passive from detractor or promotor from passive will increase the NPS score.

The decision tree suggest that Southeast Airlines Inc. should firstly focus on improving the customer experience in eating and drinking at the airport. Then the company should make different strategies for different types of travel and different age groups to get a higher NPS score.

The sentiment analysis allowed us to take a deeper dive into the specific airline partners to determine which provided their customers with a more positive experience, and which provided them with a more negative experience. By taking a deeper dive into the airline partners with negative leaning sentiment we were able to determine what factors were most frequently vocalized by these customers across both. These factors provide specific guidance towards what areas of the business need to be improved in order to have a direct impact on the overall NPS score.

Appendix A - Code Files

Data Cleaning and Outlier Analysis: The code run to give the Data Cleaning and Outlier Analysis results can be found in the file labelled

- IST687M002_Group2_FinalProject_DataCleaningCode

Linear Model: The code run to give the Linear Modeling results can be found in the file labelled

- IST687M002_Group2_FinalProject_LinearModelCode

Association Rules: The code run to give the Association Rules results can be found in the file labelled

- IST687M002_Group2_FinalProject_AssociationRulesCode

Decision Tree: The code run to give the Decision Tree and Random Forest results can be found in the file labelled

- IST687M002_Group2_FinalProject_DecisionTreeCode

Sentiment Analysis: The code run to give the Sentiment Analysis results can be found in the file labelled

- IST687M002_Group2_FinalProject_SentimentAnalysisCode

Appendix B - Bibliography

[1] IST 687, Final Project, Project Overview Document
[2] J. Saltz, J. Stanton, 2018, “Introduction to Data Science”
[3] Picture of Box Plot - https://chartio.com/resources/tutorials/what-is-a-box-plot/
[4] rpart package documentation
-https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart
[5] Random Forest Package Documentation
-https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest