IST 687 – INTRODUCTION TO DATA SCIENCE
FINAL PROJECT REPORT

Southeast Airlines: Business Strategy Analysis

Group 5

Aatish Suman

Adheesh Ajit Phadnis

Aneesh Phatak

Chaoying Lyu

Nishitha Maniganahalli Venkatesh

Contents

INTRODUCTION

Southeast Airlines wants to lower their customer churn. Customer churn is when a customer a customer chooses to stop using the services. There are several key factors that are relevant to customer churn, one of them is Net Promoter Score (NPS). Customers score the airlines on a scale of 1-10. If the score is less than 7, they are Detractors. If the score is above 8, they are Promoters. If the score is 7 or 8, they are Passive. NPS is the difference between proportion of Promoters and Detractors. Customers who are Promoters are good to keep while Detractors are problematic.

The aim of the project is to increase NPS for Southeast Airlines by determining which group of customers are likely to be Detractors.

BUSINESS RULES

1. What type of customers are likely to be promoters/detractors/passive?
2. Which origin state has the highest number of customers?
3. Which origin state has the highest proportion of detractors?
4. Is there a relationship between age groups and likelihood to recommend?
5. Which partner airlines has the highest number of customers?
6. Does gender influence likelihood to recommend?
7. Does shopping at airport have a relationship with gender?
8. How does customer reviews affect recommendation?
9. Which type of travel has more detractors?
10. Does price sensitivity affect likelihood to recommend?

DATA MUNGING

NA identification and elimination

The NA values were found in the below columns using the summary command on the dataset.

Departure.Delay.in.Minutes

Arrival.Delay.in.Minutes

Flight.time.in.minutes

Likelihood.to.recommend

freeText

```
>   summary(df)
 Destination.City   Origin.City        Airline.Status          Age           Gender          Price.Sensitivity Year.of.First.Flight
 Length:10282       Length:10282       Length:10282       Min.   :15.00   Length:10282       Min.   :0.000     Min.   :2003
 Class :character   Class :character   Class :character   1st Qu.:33.00   Class :character   1st Qu.:1.000     1st Qu.:2004
 Mode  :character   Mode  :character   Mode  :character   Median :45.00   Mode  :character   Median :1.000     Median :2007
                                                          Mean   :46.25                      Mean   :1.282     Mean   :2007
                                                          3rd Qu.:59.00                      3rd Qu.:2.000     3rd Qu.:2010
                                                          Max.   :85.00                      Max.   :4.000     Max.   :2012

 Flights.Per.Year    Loyalty        Type.of.Travel     Total.Freq.Flyer.Accts Shopping.Amount.at.Airport Eating.and.Drinking.at.Airport
 Min.   : 0.0     Min.   :-0.9762   Length:10282       Min.   : 0.0000        Min.   :  0.00             Min.   :  0.00
 1st Qu.: 9.0     1st Qu.:-0.7046   Class :character   1st Qu.: 0.0000        1st Qu.:  0.00             1st Qu.: 30.00
 Median :17.0     Median :-0.4444   Mode  :character   Median : 0.0000        Median :  0.00             Median : 60.00
 Mean   :20.1     Mean   :-0.2765                      Mean   : 0.8786        Mean   : 26.41             Mean   : 68.19
 3rd Qu.:29.0     3rd Qu.: 0.0588                      3rd Qu.: 2.0000        3rd Qu.: 30.00             3rd Qu.: 90.00
 Max.   :92.0     Max.   : 1.0000                      Max.   :10.0000        Max.   :626.00             Max.   :650.00

    Class           Day.of.Month   Flight.date        Partner.Code       Partner.Name       Origin.State       Destination.State
 Length:10282       Min.   : 1.00   Length:10282       Length:10282       Length:10282       Length:10282       Length:10282
 Class :character   1st Qu.: 8.00   Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Median :16.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
                    Mean   :15.76
                    3rd Qu.:23.00
                    Max.   :31.00

 Scheduled.Departure.Hour Departure.Delay.in.Minutes Arrival.Delay.in.Minutes Flight.cancelled   Flight.time.in.minutes Flight.Distance
 Min.   : 1.00            Min.   :  0.00             Min.   :  0.00           Length:10282       Min.   : 15            Min.   :  67.0
 1st Qu.: 9.00            1st Qu.:  0.00             1st Qu.:  0.00           Class :character   1st Qu.: 62            1st Qu.: 386.0
 Median :13.00            Median :  0.00             Median :  0.00           Mode  :character   Median : 94            Median : 643.0
 Mean   :13.04            Mean   : 15.17             Mean   : 15.51                              Mean   :116            Mean   : 832.5
 3rd Qu.:17.00            3rd Qu.: 13.00             3rd Qu.: 14.00                              3rd Qu.:147            3rd Qu.:1072.0
 Max.   :23.00            Max.   :933.00             Max.   :920.00                              Max.   :394            Max.   :3414.0
                          NA's   :208                NA's   :235                                 NA's   :235
 Likelihood.to.recommend     olong             olat           dlong             dlat          freeText
 Min.   : 1.000          Min.   :-165.39   Min.   :18.02   Min.   :-165.39   Min.   :18.02   Length:10282
 1st Qu.: 6.000          1st Qu.:-111.93   1st Qu.:33.49   1st Qu.:-111.93   1st Qu.:33.82   Class :character
 Median : 8.000          Median : -90.34   Median :37.67   Median : -90.14   Median :37.67   Mode  :character
 Mean   : 7.171          Mean   : -95.40   Mean   :37.12   Mean   : -95.24   Mean   :37.10
 3rd Qu.: 9.000          3rd Qu.: -81.61   3rd Qu.:41.07   3rd Qu.: -81.64   3rd Qu.:40.72
 Max.   :10.000          Max.   : -66.12   Max.   :71.29   Max.   : -66.12   Max.   :71.29
 NA's   :1
```

For the first 3 columns, it was found out that the value of the corresponding Flight.cancelled column was “Yes”, and it makes sense that if the flight was cancelled, there would not be a corresponding value for any of these columns, and therefore, these NA values were replaced with 0. For the Likelihood.to.recommend column, there was only 1 row which had an NA and it was replaced with the overall average of the column.

Each of the numerical columns listed below were binned into logical categories considering the range of the values for these columns. Likelihood.to.recommend was binned into *detractor, passive* and *promoter*.

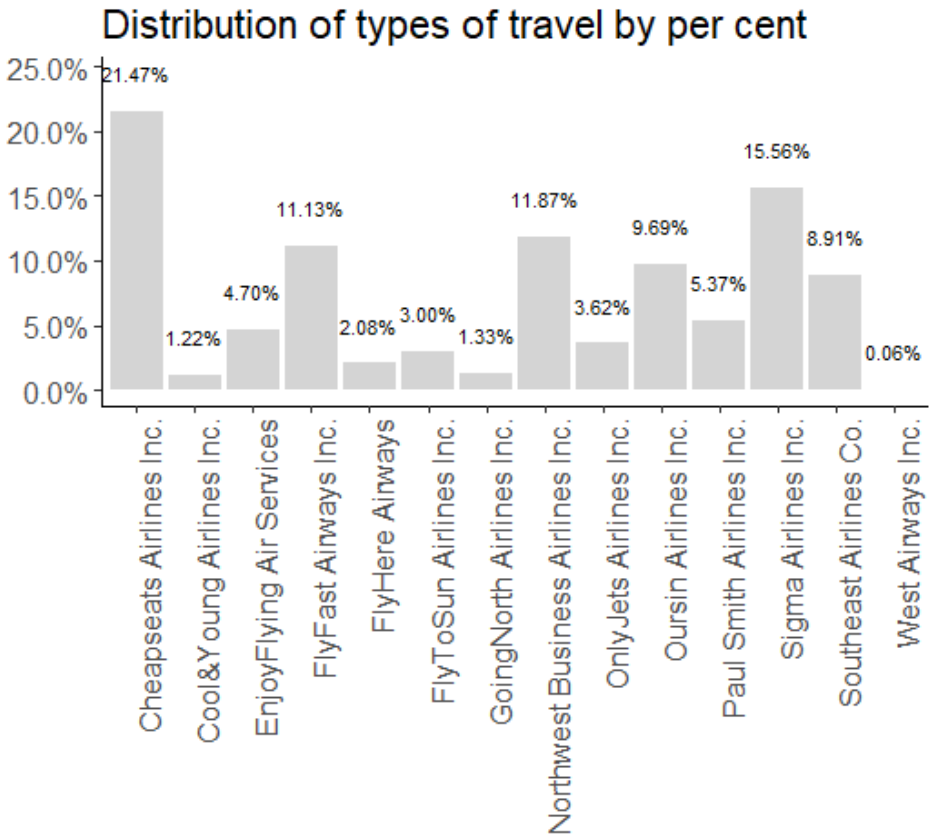
Data cleaning and binning

As the project takes four different approaches, columns are dropped and continuous variables are binned based on the flexibility for each approach.

DESCRIPTIVE STATISTICS

1. **Airlines partners frequency**

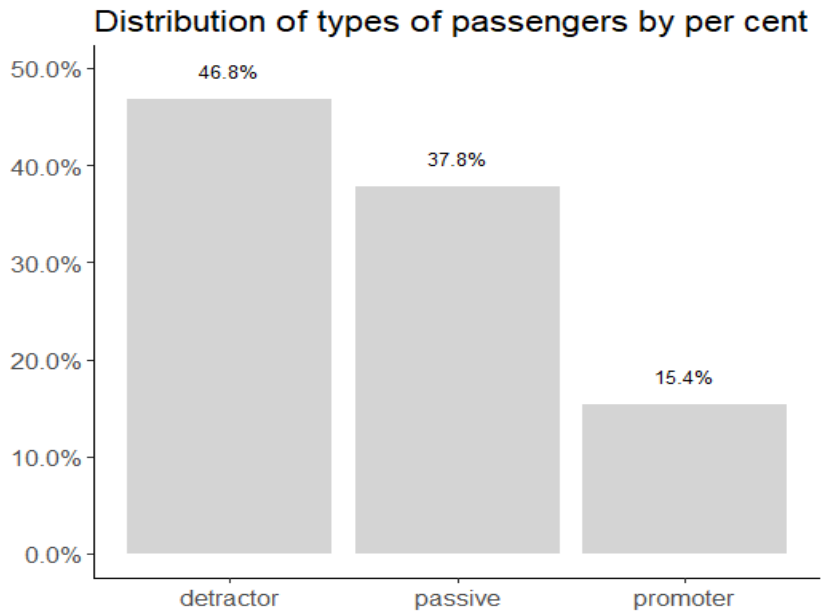
   Based on the count of flights we chose a subset of data for our first model consisting of 'Top 3 most frequent Airline Partners'. Passengers travelled in these three partners in 50% of all flights.



   It was also necessary to focus on those particular subsets of data that had the highest proportion of detractors.

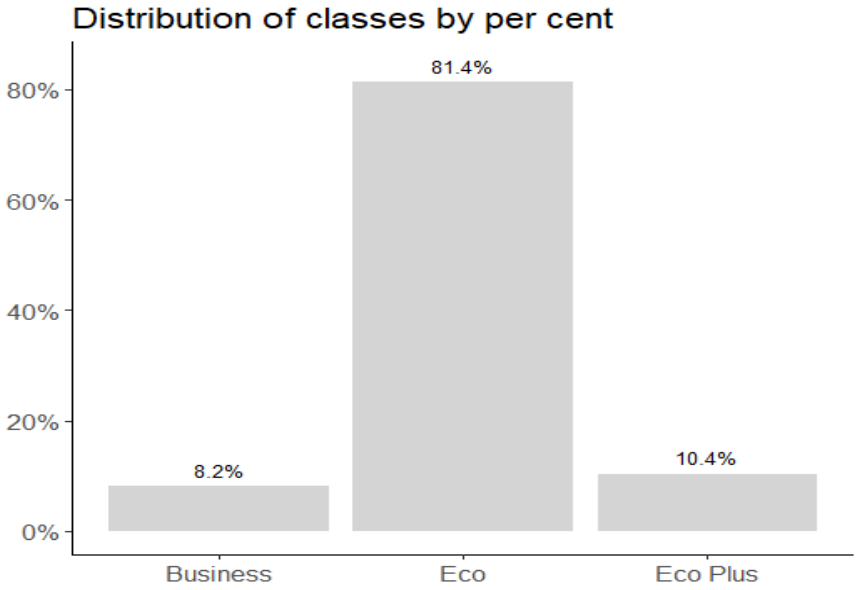
2. **Proportion of detractors overall**

   Total percentage of detractors in the entire dataset was 46.8%.

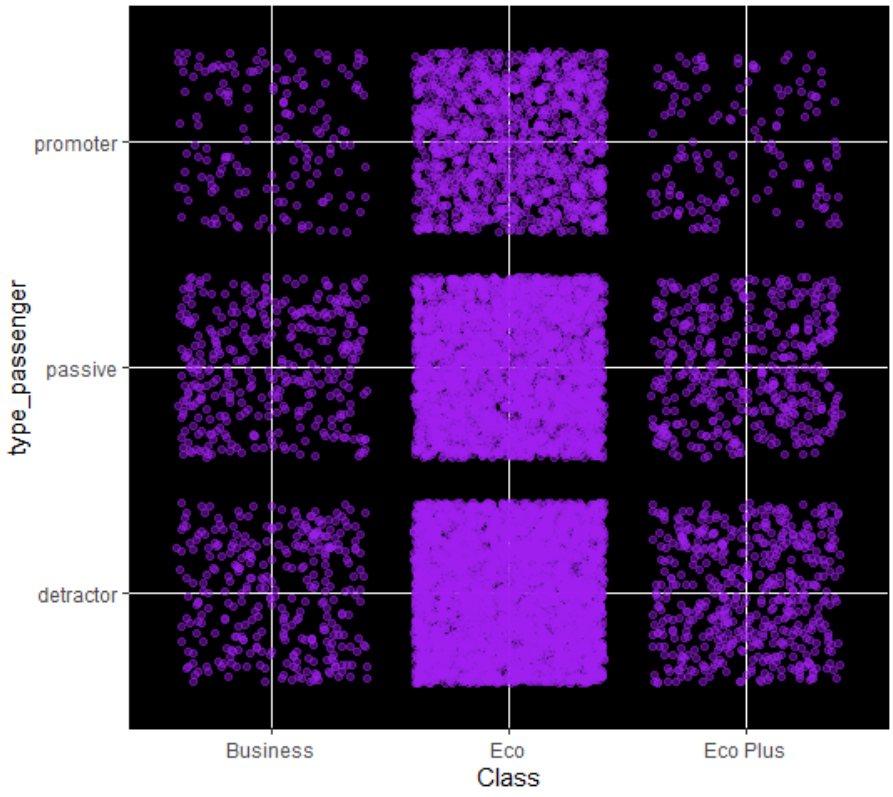


3. Class distribution

Passengers in 81.4 % of the cases flew by Eco class.

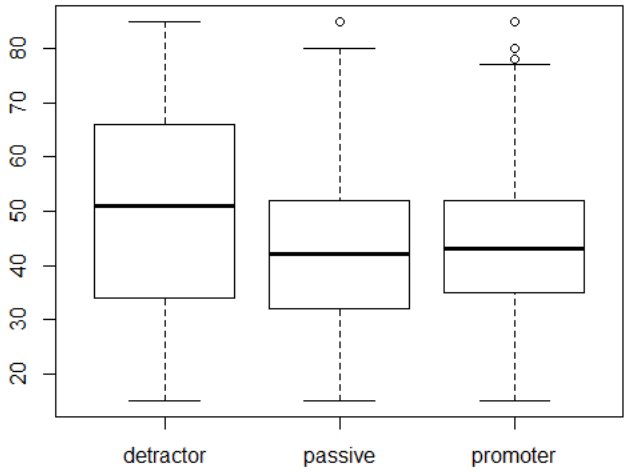


Proportion of detractors according to Class also was very high for Eco class.

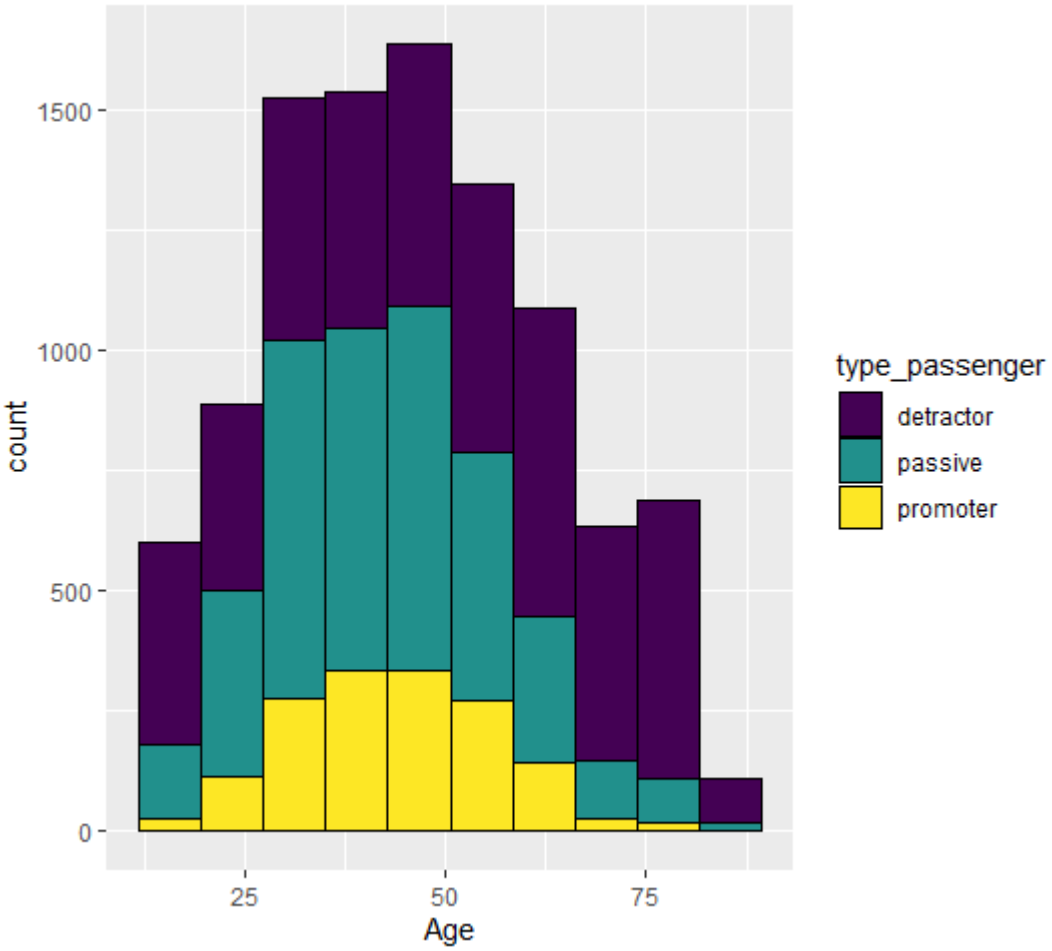


4. **Distribution of type of passenger by age**

From the box plot below, detractors seem to have median age greater than passive passengers or promoters. It is important to focus on passengers with higher age.

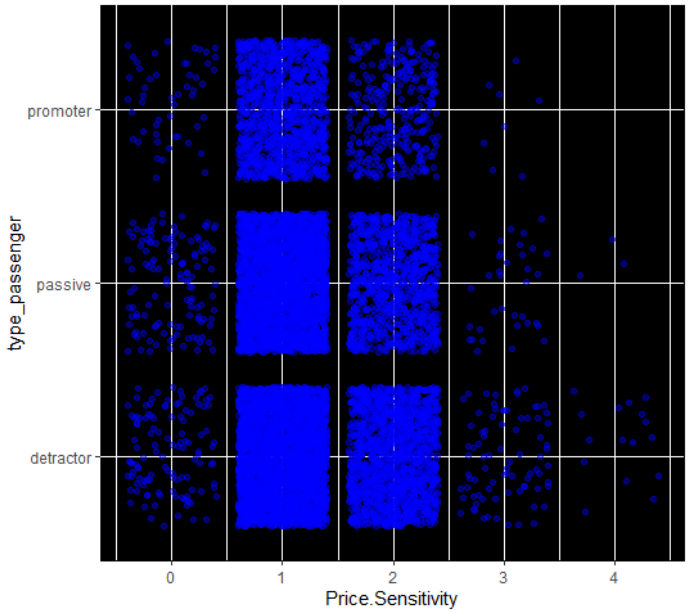


As we can see from the histogram below,majority of passengers with age greater than ~60 have a higher proportion of detractors.



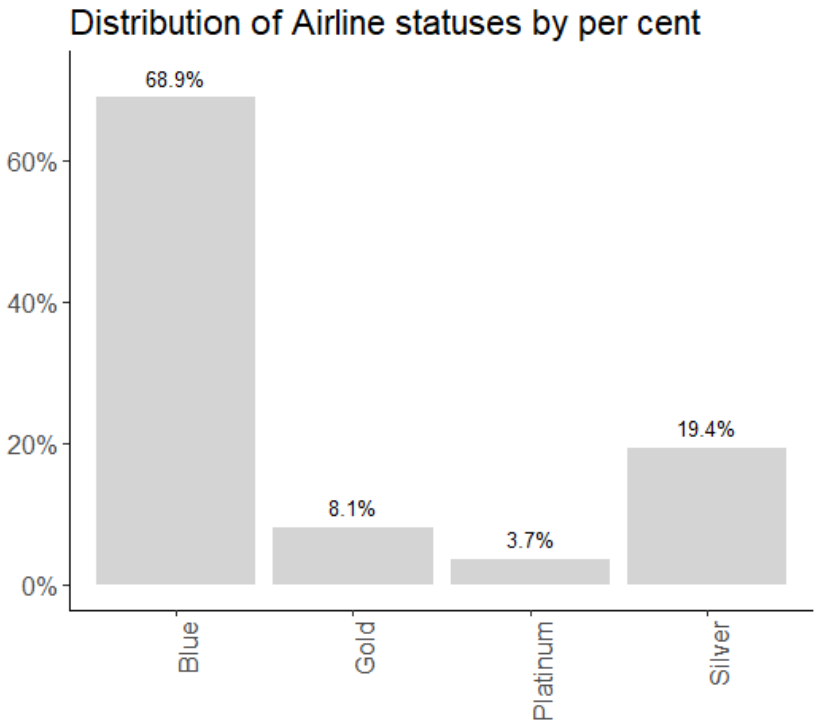
5. **Price sensitivity vs type of passenger**

When price sensitivity was 1 or 2, the proportion of detractors was high.

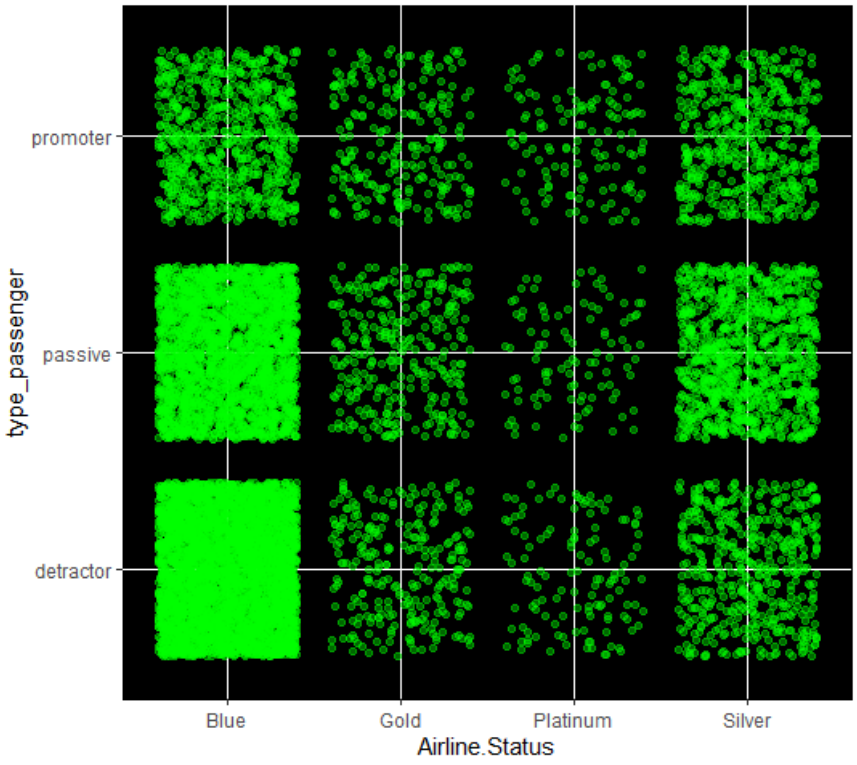


6. **Airline statuses vs type of passenger**

Airline status 'Blue' had the highest percentage of passengers.

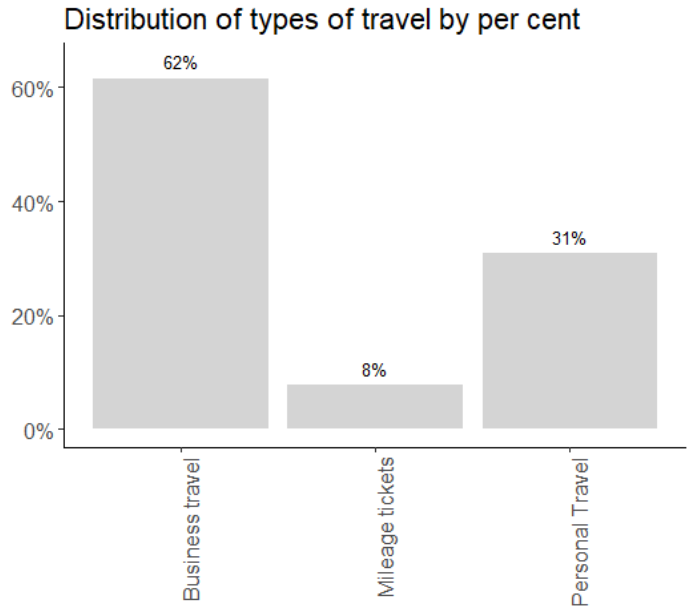


It also had the highest proportion of detractors.

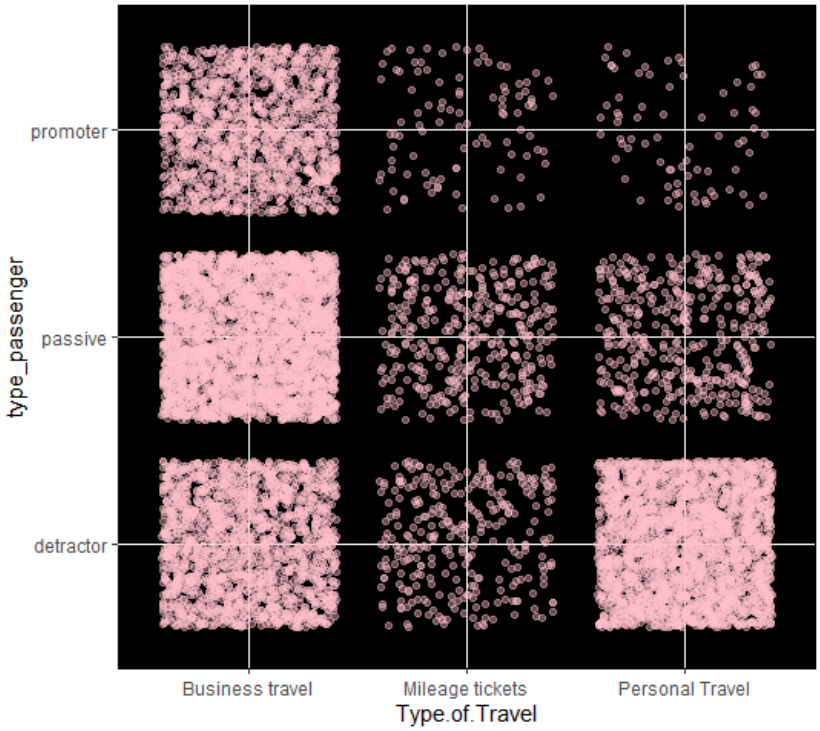


7. **Type of travel vs type of passenger**

   Of all the passengers analyzed, 62% of them travelled for business followed by 31 % for personal travel.



Even though the type of travel was personal for only 31 % of the flights, the proportion of detractors in these was the highest.



DATA MODELING

Analyzing data based on origin state

For the first approach we grouped the data by origin state and summarized the data by proportion of promoters, detractors and passive customers.

	Origin.State	Detractors	Passive	Promoters	Total
1	texas	0.62597114	0.3362930	0.03773585	901
2	connecticut	0.44680851	0.2978723	0.25531915	47
3	south carolina	0.41818182	0.2545455	0.32727273	55
4	kansas	0.38095238	0.1428571	0.47619048	21
5	north dakota	0.37037037	0.3333333	0.29629630	27
6	rhode island	0.36842105	0.2631579	0.36842105	19
7	hawaii	0.35897436	0.2564103	0.38461538	39
8	tennessee	0.35593220	0.3050847	0.33898305	59
9	indiana	0.34090909	0.2500000	0.40909091	88
10	virginia	0.34083601	0.3118971	0.34726688	311
11	ohio	0.33838384	0.3232323	0.33838384	198
12	new jersey	0.33333333	0.1666667	0.50000000	6
13	massachusetts	0.32000000	0.2550000	0.42500000	200
14	illinois	0.31457801	0.3299233	0.35549872	782
15	wyoming	0.30769231	0.4230769	0.34615385	26
16	oklahoma	0.30681818	0.2272727	0.46590909	88
17	nevada	0.30372493	0.3070043	0.39838080	349

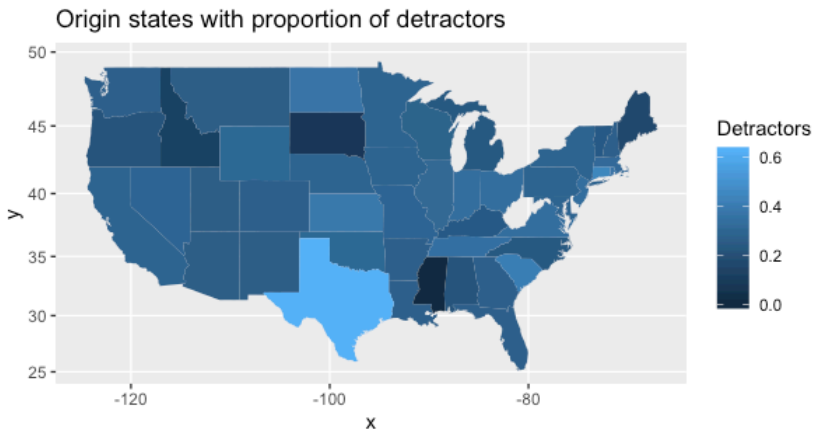
Showing 1 to 18 of 49 entries, 5 total columns

It was found that Texas state had maximum number of customers. Approximately 60% of them were detractors.

Data engineering

Price sensitivity with 0 and 1 is binned to low, 2 and 3 to moderate, and 4 and 5 as high.

Columns like destination city, origin city, longitude and latitude, year and day are omitted.



Upon digging deeper into the Texas state data and running Apriori algorithm, the following rule sets were generated.

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[13]	{Airline.Status=Blue,Type.of.Travel=Personal Travel}	{Likelihood.to.recommend=detractor}	0.246	0.965	1.542	222.000
[85]	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Flight.cancelled=No}	{Likelihood.to.recommend=detractor}	0.240	0.969	1.547	216.000
[14]	{Type.of.Travel=Personal Travel,Class=Eco}	{Likelihood.to.recommend=detractor}	0.232	0.925	1.477	209.000
[86]	{Type.of.Travel=Personal Travel,Class=Eco,Flight.cancelled=No}	{Likelihood.to.recommend=detractor}	0.224	0.927	1.480	202.000
[12]	{Gender=Female,Type.of.Travel=Personal Travel}	{Likelihood.to.recommend=detractor}	0.205	0.920	1.470	185.000
[80]	{Gender=Female,Type.of.Travel=Personal Travel,Flight.cancelled=No}	{Likelihood.to.recommend=detractor}	0.201	0.928	1.483	181.000
[83]	{Price.Sensitivity=low,Type.of.Travel=Personal Travel,Flight.cancelled=No}	{Likelihood.to.recommend=detractor}	0.198	0.904	1.443	178.000

The results of Texas state data indicate that customers with airline status Blue and Personal travel are major detractors. Female travelers are also the common parameter among the strongest rules.

Advanced model 1 (Top 3 airline partners)

After doing some exploratory analysis, it was found that the Cheapseats Airlines Inc., Sigma Airlines Inc. and Northwest Business Airlines Inc. have the highest number of instances. This model focuses on these three airline partners. After data cleaning and feature engineering, a clean subset of this data is created.

Model 1: Association rules mining using Apriori Algorithm

The target variable, Likelihood to recommend is binned broadly into two categories.

1. Customer ratings < 8 : Detractors
2. Customer Ratings >= : Promoters

Apriori algorithm is used to find out rules for detractors and promoters.

Using these rules, most frequent variables occurring in the rules are identified and only those variables are used to build the next model.

Model 2: Support Vector Machines

It was observed in the rules generated in model 1 that most detractors happen to be in Cheapseats airlines Inc. This airline partner name is associated with detractors in the rules generated for detractors which signifies that the airline should focus specifically on this airline partner. Hence, using the variables identified in model 1 and instances for Cheapseats airlines, SVM was used to classify the detractors and promoters.

Data Cleaning and Feature Engineering –

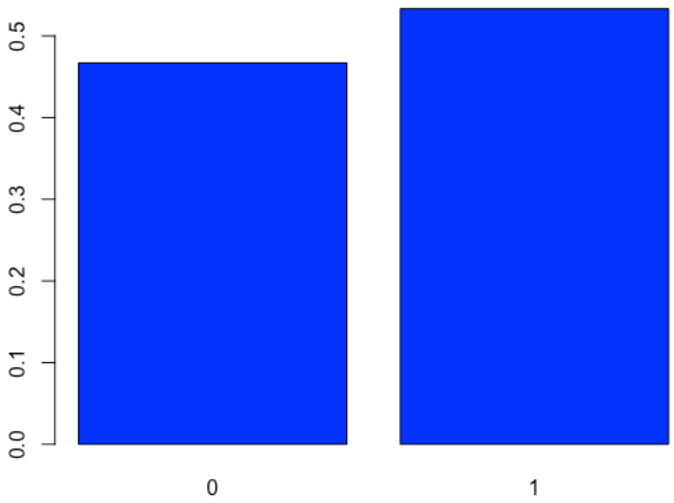
1. Dropping columns that occur repetitively/ irrelevant intuitively
    a. Free Text – Very less data and irrelevant to modeling technique
    b. Year of Flight – Cannot see intuitive contribution to model
    c. Flight Date - Cannot see intuitive contribution to model
    d. Partner Code – Because Partner Name is used, repetitive
    e. Arrival delay in minutes – Departure delay is used
    f. Latitude, longitude coordinates – Irrelevant to modeling approach
    g. Destination City – Destination state is used
    h. Origin City – Destination state is used

2. Binning the continuous variables in the data set

a. Age - Binning the Age column into 3 categories viz. 15-35, 35-60 and >60 named as ‘young’, ‘middle aged’ and ‘old’, respectively.
b. Flights per year - Binning the flights per year column into 3 categories viz. '0-30','30-60' and '60-100'
c. Loyalty - Binning the loyalty column with 0.2 interval
d. Shopping amount at airport - Binning the Shopping amount at airport column with 10$ interval
e. Eating amount at airport - Binning the Eating amount at airport column with 20$ interval
f. Departure Delay in minutes - Binning the Departure Delay in minutes column with 20 mins interval
g. Flight time in minutes – Binning the Flight time in minutes with 100 mins interval
h. Flight distance – Binning the Flight distance column with 500 intervals
i. Factorizing the variables that are factors but stored as numeric

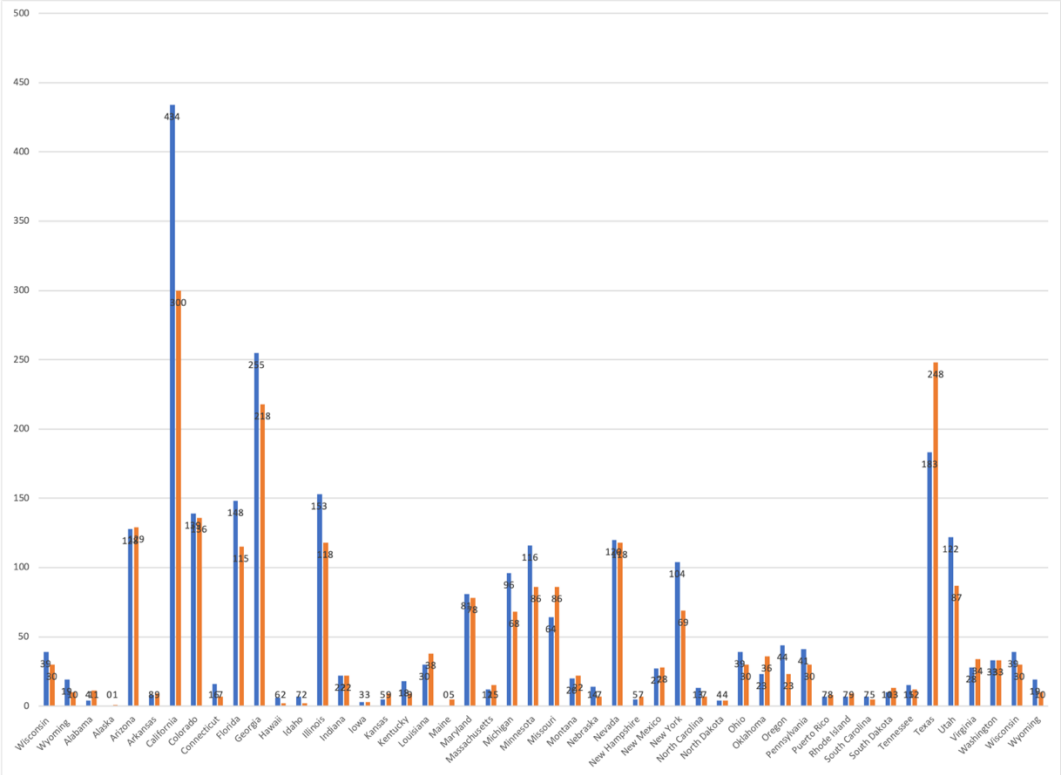
Preliminary analysis on the cleaned data –

1. Distribution of Likelihood to recommend after binning

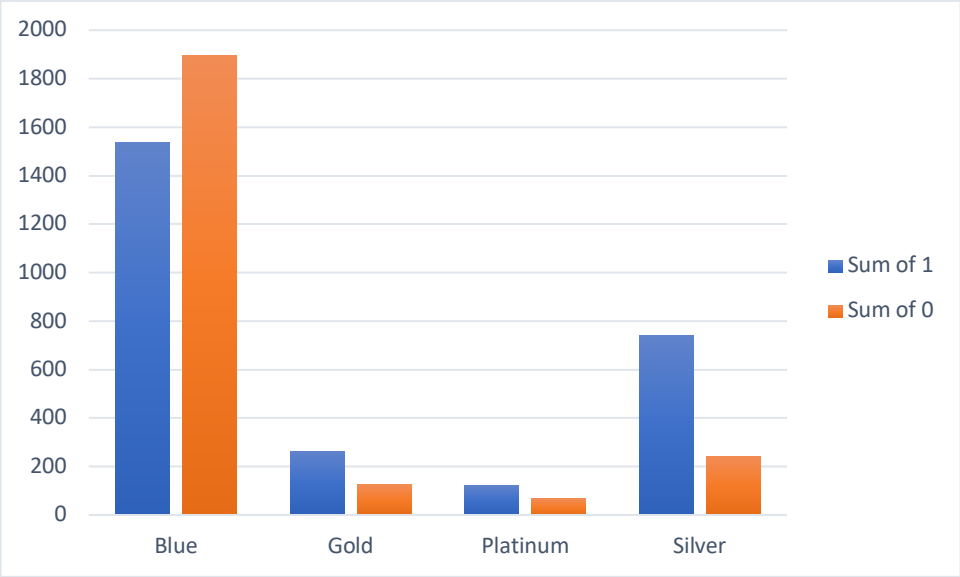


Detractors have significant proportion in the dataset (46%)

2. State wise detractors and promoters

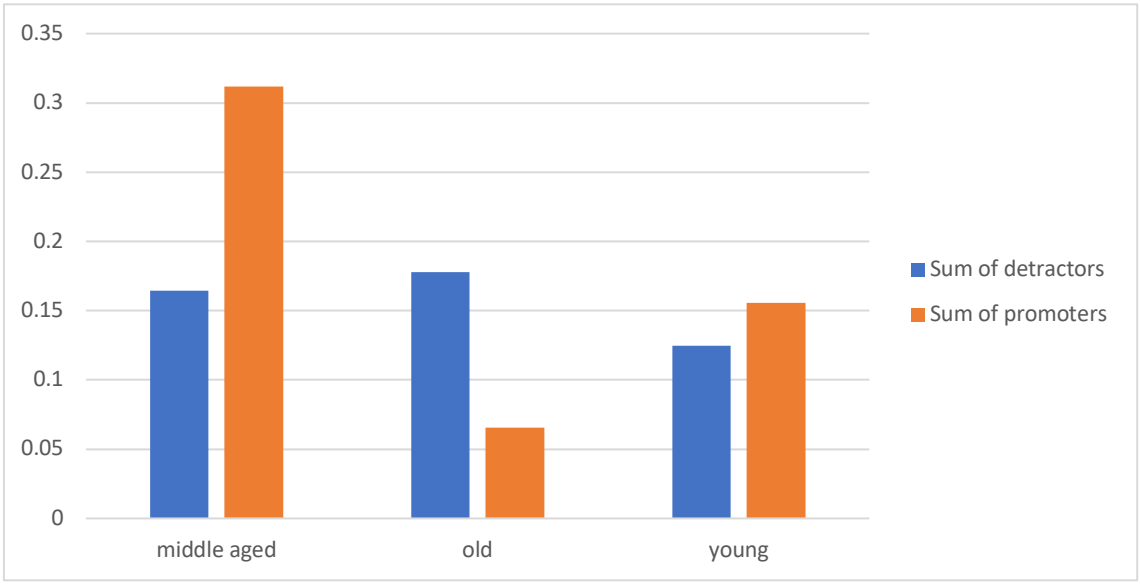


3. Distribution of detractors and promoters based on airlines status



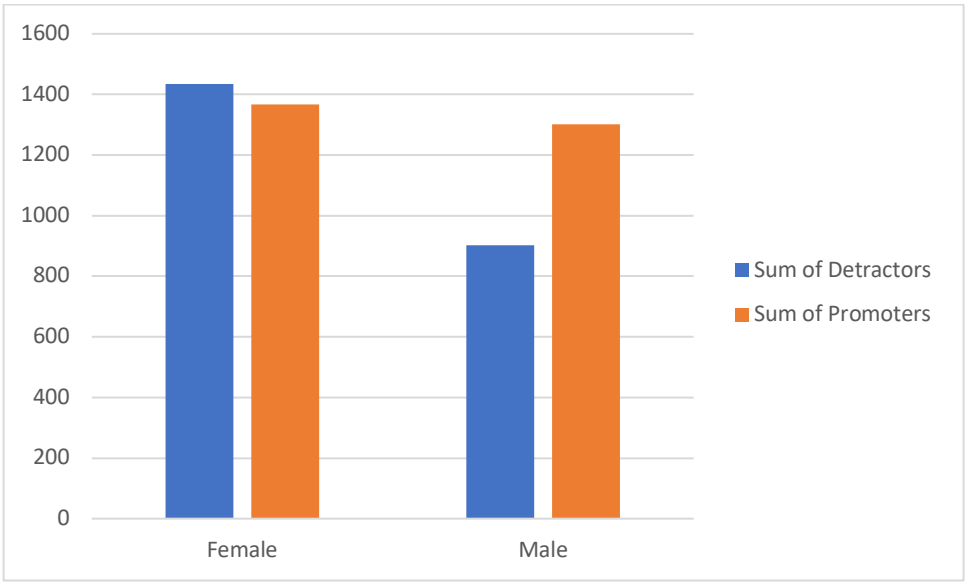
Airline status blue has the most detractors and most customers belong to blue status

4. Age wise distribution of detractors and promoters



Highest proportion of detractors are with age category more than 60

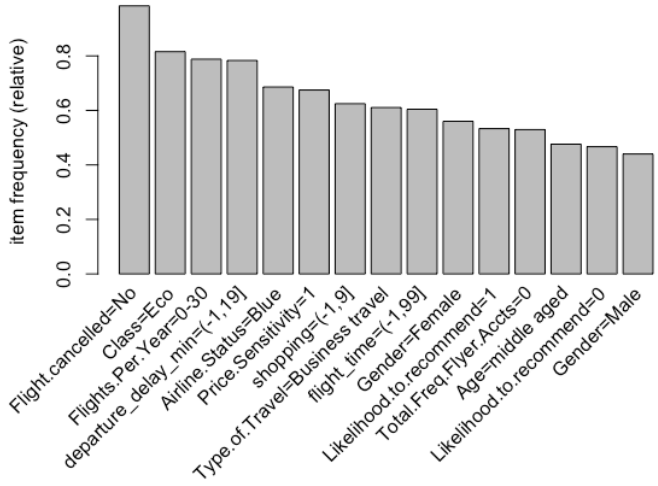
5. Gender wise distribution of detractors and promoters



Maximum detractors are female customers

Model 1: Association rule mining

1. #creating transaction matrix from the final cleaned data set
   my_data <- as(data_top3v2,'transactions')

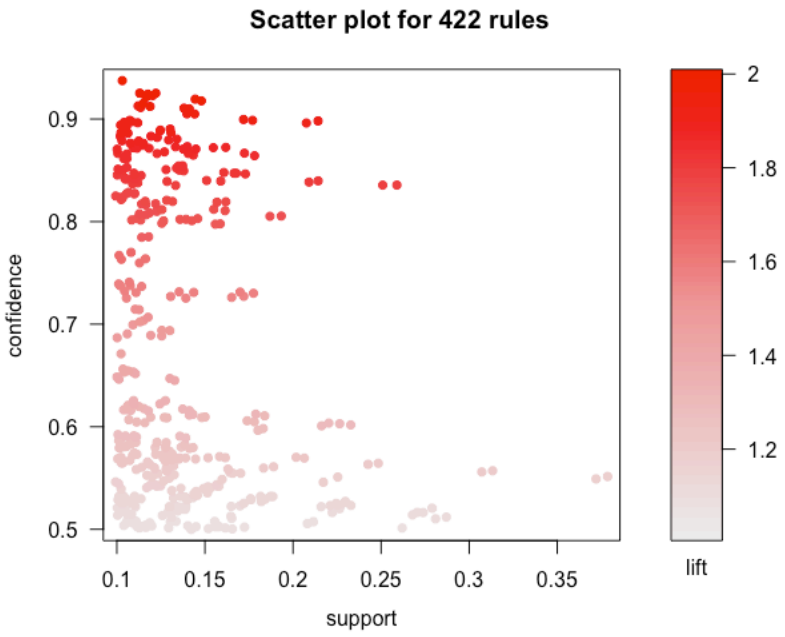


2. Building rules using apriori algorithm on my_data with support 0.1, confidence 0.5 and maxlength of rules as 20. Here, RHS is detractors. meaning the rules will specify what characteristics define detractors.

   ruleset <- apriori(my_data, parameter = list(support = 0.1,confidence = 0.5, maxlen = 20), appearance = list(default = 'lhs', rhs = ("Likelihood.to.recommend=0")))

```
> ruleset <- apriori(my_data, parameter = list(support = 0.1,confidence = 0.5, maxlen = 20), appearance = list(default = 'lhs', rhs = ("Likelihood.to.recommend=0")))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5     0.1      1     20  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 500

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[270 item(s), 5006 transaction(s)] done [0.01s].
sorting and recoding items ... [43 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.06s].
writing ... [422 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Ruleset for 422 rules



After inspecting the ruleset, it can be seen that there are 422 rules generated. Since it is manually tedious to analyze 422 rules, we have narrowed down the rules with lift greater than 1.8.

Good rules for detractors with lift > 1.8 –

Show 10 entries Search:

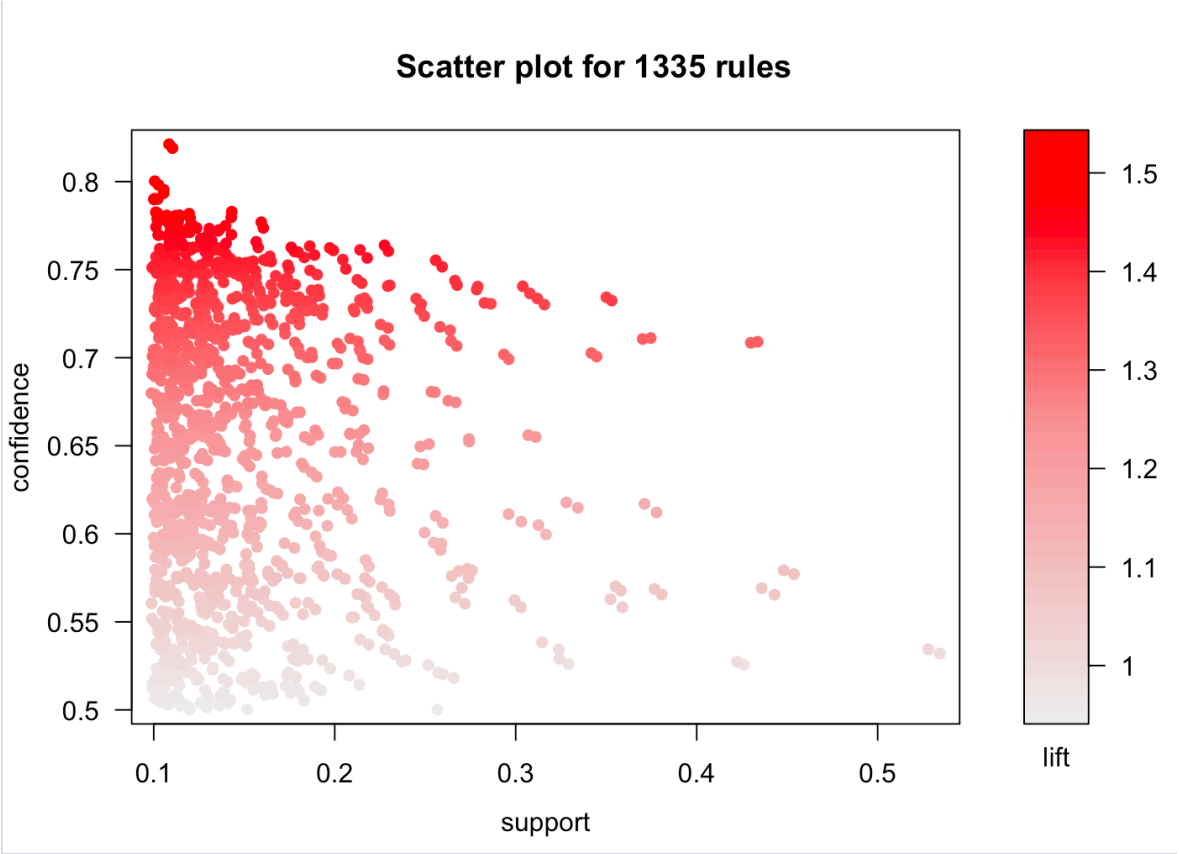
	lhs	rhs	support	confidence	lift	count
1	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Partner.Name=Cheapseats Airlines Inc.}	{Likelihood.to.recommend=0}	0.10207750699161	0.935897435897436	2.0047507762527	511
2	{Airline.Status=Blue,Age=old,Type.of.Travel=Personal Travel,Flight.cancelled=No}	{Likelihood.to.recommend=0}	0.118457850579305	0.925117004680187	1.98165841909671	593
3	{Airline.Status=Blue,Age=old,Type.of.Travel=Personal Travel,Total.Freq.Flyer.Accts=0,Flight.cancelled=No}	{Likelihood.to.recommend=0}	0.11286456252497	0.924713584288052	1.98079426741377	565
4	{Airline.Status=Blue,Age=old,Type.of.Travel=Personal Travel}	{Likelihood.to.recommend=0}	0.122053535757091	0.924357034795764	1.98003051612648	611
5	{Airline.Status=Blue,Age=old,Type.of.Travel=Personal Travel,Total.Freq.Flyer.Accts=0}	{Likelihood.to.recommend=0}	0.116460247702757	0.923930269413629	1.97911635801653	583
6	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Total.Freq.Flyer.Accts=0,Class=Eco,Flight.cancelled=No}	{Likelihood.to.recommend=0}	0.116260487415102	0.922345483359746	1.97572164728237	582
7	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Total.Freq.Flyer.Accts=0,Class=Eco}	{Likelihood.to.recommend=0}	0.119256891729924	0.921296296296296	1.97347422304632	597
8	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Total.Freq.Flyer.Accts=0,Flight.cancelled=No}	{Likelihood.to.recommend=0}	0.144027167399121	0.918471337579618	1.96742298499083	721

3. Building rules using apriori algorithm on my_data with support 0.1, confidence 0.5 and maxlength of rules as 20. Here, RHS is PROMOTERS. meaning the rules will specify what characteristics define PROMOTERS

ruleset1 <- apriori(my_data, parameter = list(support = 0.1,confidence = 0.5, maxlen = 20), appearance = list(default = 'lhs', rhs = ("Likelihood.to.recommend=1")))

```
> ruleset1 <- apriori(my_data, parameter = list(support = 0.1,confidence = 0.5, maxlen = 20), appearance = list(default = 'lhs', rhs = ("Likelihood.to.recommend=1")))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5     0.1      1     20  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 500

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[270 item(s), 5006 transaction(s)] done [0.01s].
sorting and recoding items ... [43 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.06s].
writing ... [1335 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

1335 rules were generated for promoters



Good rules with lift > 1.4

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[1]	{Airline.Status=Silver}	{Likelihood.to.recommend=1}	0.148	0.751	1.409	740.000
[2]	{Airline.Status=Silver,Type.of.Travel=Business travel}	{Likelihood.to.recommend=1}	0.110	0.818	1.535	550.000
[3]	{Airline.Status=Silver,Price.Sensitivity=1}	{Likelihood.to.recommend=1}	0.111	0.774	1.452	558.000
[4]	{Airline.Status=Silver,departure_delay_min=(-1,19]}	{Likelihood.to.recommend=1}	0.120	0.778	1.460	600.000
[5]	{Airline.Status=Silver,Flights.Per.Year=0-30}	{Likelihood.to.recommend=1}	0.127	0.766	1.437	638.000
[6]	{Airline.Status=Silver,Flight.cancelled=No}	{Likelihood.to.recommend=1}	0.147	0.755	1.415	735.000
[7]	{Type.of.Travel=Business travel,Partner.Name=Northwest Business Airlines Inc.}	{Likelihood.to.recommend=1}	0.112	0.754	1.415	562.000
[8]	{Airline.Status=Silver,Type.of.Travel=Business travel,Flight.cancelled=No}	{Likelihood.to.recommend=1}	0.109	0.821	1.540	547.000
[9]	{Airline.Status=Silver,Price.Sensitivity=1,Flight.cancelled=No}	{Likelihood.to.recommend=1}	0.110	0.777	1.457	553.000
[10]	{Airline.Status=Silver,Flights.Per.Year=0-30,departure_delay_min=(-1,19]}	{Likelihood.to.recommend=1}	0.103	0.797	1.495	515.000

Showing 1 to 10 of 188 entries Previous 1 2 3 4 5 … 19 Next

Model 2: Support Vector Machine

Show 10 entries Search:

	lhs	rhs	support	confidence	lift	count
1	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Partner.Name=Cheapseats Airlines Inc.}	{Likelihood.to.recommend=0}	0.10207750699161	0.935897435897436	2.0047507762527	511

It can be seen that the highest lift is for the rule in the screenshot above. Cheapseats Airlines is the airline with highest customers and also associated with detractors as shown in the exploratory data analysis. The support for this claim is again bolstered through the association rules mining.

Hence, we will now narrow down the data set to Cheapseats Airlines and the variables that occur most frequently in the rules generated for detractors and promoters.

After analyzing the rules, following variables are selected.

1. Type of travel
2. Airline Status
3. Gender
4. Class
5. Flight Cancelled

6. Departure Delay
7. Flight Time
8. Total Freq Flyer Acc
9. Age
10. Flight Time
11. Flight Distance
12. Shopping
13. Flights per year

Converting categorical data to numeric data using one hot encoding –

The data set that we have created contains only categorical variables since ARM required all categorical variables. SVM requires the independent variables to be numeric. We have decided to create dummy variables for all the 13 variables selected using one hot encoding.

```
#Librarying Mltools package which contains one_hot function
library(mltools)
library(data.table)

cheapseatsv1 <- one_hot(as.data.table(cheapseats))
```

Train test Split –

Training - 70%

Testing 30%

Model Tuning –

Using svm.tune function from package e1071, we tune the model to get the appropriate cost parameter.

```
#Tuning the model to get the optimum cost level
library(e1071)
library(kernlab)
tune <- tune.svm(target~., data=training, cost=c(0.001, 0.01, 0.1, 1,5,10))
summary(tune)
```

Tuning Output –

```
> summary(tune)

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
  0.1

- best performance: 0.268428
```

We can see that from the range provided, the model performs the best at cost= 0.1

Model Characteristics –

Data	Training
Kernel	Radial(rbfdot)
kpar	automatic
Cost	0.1
Cross	3
Probability model	True
Function used	ksvm

Model output:

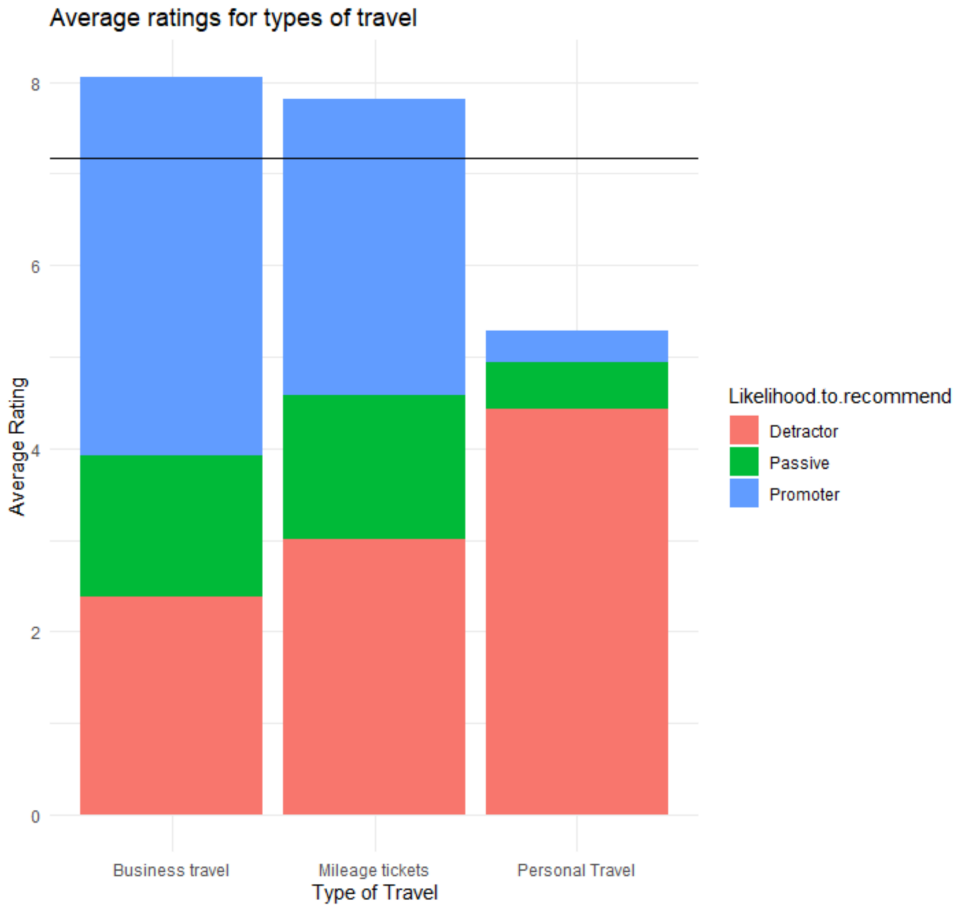
```
> svmOutput
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 0.1

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.09375

Number of Support Vectors : 1146

Objective Function Value : -99.139
Training error : 0.264495
Cross validation error : 0.269683
Probability model included.
```

Testing the model on test data –

```
> confusionMatrix(svmPred,testing$target)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 193  35
         1 148 281

               Accuracy : 0.7215
                 95% CI : (0.6855, 0.7554)
    No Information Rate : 0.519
    P-Value [Acc > NIR] : < 2.2e-16
```

Advanced model 2 (Personal travel)

Exploratory analysis



The figure shows the distribution of the likelihood to recommend values for the different types of travel. As can be seen from the plot, the average rating of passengers with type of travel as "Personal Travel" (5.28) is considerably lower than the overall average (7.17, represented by the horizontal line in the plot) as well as the average values of the other types (8.05 & 7.81). Moreover, about 84% of the passengers of this category tend to be detractors. As a result, based on these statistics, one of the rationales behind choosing this dataset for in-depth analysis is based on the fact that if such a large percentage of passengers of this category have this one feature in common (that they are detractors), it is more than likely that they have some other features in common also which directly affect their ratings.

The data was cleaned to remove NA values and the columns not needed, and a combination of association rule mining and classification modeling was used to further validate these inferences.

Data munging and feature addition

Numerical columns –

Age	Price.Sensitivity	Year.Of.First.Flight
Flights.Per.Year	Loyalty	Total.Freq.Flyer.Accts
Shopping.Amount.at.Airport	Eating.and.Drinking.at.Airport	Day.of.Month
Scheduled.Departure.Hour	Departure.Delay.in.Minutes	Arrival.Delay.in.Minutes
Flight.time.in.minutes	Flight.Distance	olong
olat	dlong	dlat

```
  # Age (15-85)
  data = data %>%
    mutate(Age = cut(Age, breaks = c(min(Age), 25, 35, 45, 55, 65, 75,
max(Age)), labels = c("<25", "25-35", "35-45", "45-55", "55-65", "65-75",
">75"), include.lowest = TRUE))

  # Year.Of.First.Flight (2003-2012)
  data = data %>%
    mutate(Year.of.First.Flight = cut(Year.of.First.Flight, breaks =
c(min(Year.of.First.Flight), 2006, 2009, max(Year.of.First.Flight)), labels =
c("before 2006", "2006-09", "after 2009"), include.lowest = TRUE))

  # Flights.Per.Year (0-100)
  data = data %>%
    mutate(Flights.Per.Year = cut(Flights.Per.Year, breaks =
c(min(Flights.Per.Year), 40, max(Flights.Per.Year)), labels = c("<40",
">40"), include.lowest = TRUE))

  # Loyalty (-1-1)
  data = data %>%
    mutate(Loyalty = cut(Loyalty, breaks = c(min(Loyalty), -0.5, 0, 0.5,
max(Loyalty)), labels = c("<-0.5", "-0.5-0", "0-0.5", ">0.5"), include.lowest
= TRUE))

  # Shopping.Amount.at.Airport (0-600)
  data = data %>%
    mutate(Shopping.Amount.at.Airport = cut(Shopping.Amount.at.Airport,
breaks = c(min(Shopping.Amount.at.Airport), 100, 200, 300, 400, 500,
max(Shopping.Amount.at.Airport)), labels = c("<100", "100-200", "200-300",
"300-400", "400-500", ">500"), include.lowest = TRUE))

  # Eating.and.Drinking.at.Airport (0-500)
  data = data %>%
    mutate(Eating.and.Drinking.at.Airport = cut(Eating.and.Drinking.at.Air-
port, breaks = c(min(Eating.and.Drinking.at.Airport), 100, 200, 300, 400,
max(Eating.and.Drinking.at.Airport)), labels = c("<100", "100-200", "200-
300", "300-400", ">400"), include.lowest = TRUE))

  # Day.of.Month (1-31)
  data = data %>%
    mutate(Day.of.Month = cut(Day.of.Month, breaks = c(min(Day.of.Month), 15,
max(Day.of.Month)), labels = c("before 15th", "after 15th"), include.lowest =
TRUE))
```

```
  # Scheduled.Departure.Hour (0-24)
  data = data %>%
    mutate(Scheduled.Departure.Hour = cut(Scheduled.Departure.Hour, breaks =
c(min(Scheduled.Departure.Hour), 12, max(Scheduled.Departure.Hour)), labels =
c("before 12pm", "after 12pm"), include.lowest = TRUE))

  # Departure.Delay.in.Minutes (0-550)
  data = data %>%
    mutate(Departure.Delay.in.Minutes = cut(Departure.Delay.in.Minutes,
breaks = c(min(Departure.Delay.in.Minutes), 100, 200, 300, 400,
max(Departure.Delay.in.Minutes)), labels = c("<100", "100-200", "200-300",
"300-400", ">400"), include.lowest = TRUE))

  # Arrival.Delay.in.Minutes (0-650)
  data = data %>%
    mutate(Arrival.Delay.in.Minutes = cut(Arrival.Delay.in.Minutes, breaks =
c(min(Arrival.Delay.in.Minutes), 100, 200, 300, 400, 500,
max(Arrival.Delay.in.Minutes)), labels = c("<100", "100-200", "200-300",
"300-400", "400-500", ">500"), include.lowest = TRUE))

  # Flight.time.in.minutes (0-400)
  data = data %>%
    mutate(Flight.time.in.minutes = cut(Flight.time.in.minutes, breaks =
c(min(Flight.time.in.minutes), 100, 200, 300, max(Flight.time.in.minutes)),
labels = c("<100", "100-200", "200-300", ">300"), include.lowest = TRUE))

  # Flight.Distance (0-3000)
  data = data %>%
    mutate(Flight.Distance = cut(Flight.Distance, breaks =
c(min(Flight.Distance), 1000, 2000, max(Flight.Distance)), labels =
c("<1000", "1000-2000", ">2000"), include.lowest = TRUE))

  # olong (-170 - -60)
  data = data %>%
    mutate(olong = cut(olong, breaks = c(min(olong), -145, -120, -95,
max(olong)), labels = c("<-145", "-145 - -120", "-120 - -95", ">-95"),
include.lowest = TRUE))

  # olat (15-75)
  data = data %>%
    mutate(olat = cut(olat, breaks = c(min(olat), 30, 45, 60, max(olat)),
labels = c("<30", "30-45", "45-60", ">60"), include.lowest = TRUE))

  # dlong (-170 - -60)
  data = data %>%
    mutate(dlong = cut(dlong, breaks = c(min(dlong), -145, -120, -95,
max(dlong)), labels = c("<-145", "-145 - -120", "-120 - -95", ">-95"),
include.lowest = TRUE))

  # dlat (15-65)
  data = data %>%
    mutate(dlat = cut(dlat, breaks = c(min(dlat), 30, 45, 60, max(dlat)),
labels = c("<30", "30-45", "45-60", ">60"), include.lowest = TRUE))

  # Likelihood.to.recommend (1-10)
  data = data %>%
    mutate(Likelihood.to.recommend = cut(Likelihood.to.recommend, breaks =
c(min(Likelihood.to.recommend), 7, 8, max(Likelihood.to.recommend)), labels =
c("Detractor", "Passive", "Promoter"), include.lowest = TRUE))
```

The below columns were then removed from the dataset as a basic linear model on the data did not show a strong dependency on the Likelihood.to.recommend for these columns.

Destination.City

Origin.City

Partner.Name

Flight.date

freeText

Association rules mining

2 rulesets were considered –

rhs = "Likelihood.to.recommend=Promoter", support = 0.005, confidence = 0.5, lift = 2.35 (25 rules)

rhs = "Likelihood.to.recommend=Detractor", support = 0.05, confidence = 0.8, lift = 1.99 (13 rules)

```
rulesetPromoters <- apriori(dfTnx,
                            parameter=list(support=0.005,confidence=0.5),
                            appearance = list(default="lhs",
rhs=("Likelihood.to.recommend=Promoter")))

rulesetDetractors <- apriori(dfTnx,
                             parameter=list(support=0.05,confidence=0.8),
                             appearance = list(default="lhs",
rhs=("Likelihood.to.recommend=Detractor")))
arules::inspect(rulesetPromoters[quality(rulesetPromoters)$lift > 2.35])
arules::inspect(rulesetDetractors[quality(rulesetDetractors)$lift > 1.99])
```

Promoters (top 6 rules) –

	LHS	RHS	support	confidence	lift	count
	All	All	All	All		All
[1]	{Airline.Status=Silver,Type.of.Travel=Mileage tickets,Departure.Delay.in.Minutes=<100,olong=>-95}	{Likelihood.to.recommend=Promoter}	0.005	0.885	2.424	54.000
[2]	{Airline.Status=Silver,Type.of.Travel=Mileage tickets,Arrival.Delay.in.Minutes=<100,olong=>-95}	{Likelihood.to.recommend=Promoter}	0.005	0.871	2.385	54.000
[3]	{Airline.Status=Gold,Type.of.Travel=Business travel,Partner.Code=OO,Flight.time.in.minutes=<100}	{Likelihood.to.recommend=Promoter}	0.005	0.869	2.379	53.000
[4]	{Airline.Status=Gold,Type.of.Travel=Business travel,Partner.Code=OO,Flight.Distance=<1000}	{Likelihood.to.recommend=Promoter}	0.006	0.866	2.370	58.000
[5]	{Airline.Status=Silver,Type.of.Travel=Mileage tickets,Departure.Delay.in.Minutes= <100,Arrival.Delay.in.Minutes=<100,olong=>-95}	{Likelihood.to.recommend=Promoter}	0.005	0.883	2.419	53.000
[6]	{Airline.Status=Silver,Type.of.Travel=Mileage tickets,Departure.Delay.in.Minutes= <100,Flight.cancelled=No,olong=>-95}	{Likelihood.to.recommend=Promoter}	0.005	0.883	2.419	53.000

Detractors (top 6 rules) –

	LHS	RHS	support	confidence	lift	count
	All	All	All	All		All
[1]	{Airline.Status=Blue,Flights.Per.Year=>40,Type.of.Travel=Personal Travel}	{Likelihood.to.recommend=Detractor}	0.050	0.961	2.034	515.000
[2]	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Total.Freq.Flyer.Accts= 0,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.051	0.944	1.998	524.000
[3]	{Airline.Status=Blue,Gender=Female,Type.of.Travel=Personal Travel,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.050	0.942	1.993	518.000
[4]	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Flight.Distance=<1000,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.054	0.941	1.992	559.000
[5]	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Flight.cancelled=No,Flight.Distance=<1000,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.053	0.943	1.995	544.000
[6]	{Airline.Status=Blue,Type.of.Travel=Personal Travel,Shopping.Amount.at.Airport= <100,Flight.cancelled=No,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.067	0.941	1.992	688.000

Inferences –

- Type of travel and airline status are strongly related.
- Passengers with mileage tickets on airlines with status as “Silver” are found to be promoters.
- Passengers with type of travel as personal on airlines with status as “Blue” are found to be detractors.

The columns with high association with each other were considered for classification modeling – Airline.Status

Type.of.Travel

Eating.and.Drinking.at.Airport

Departure.Delay.in.Minutes

Flights.Per.Year

Price.Sensitivity

olong

dlat

Total.Freq.Flyer.Accts

Classification model

The columns were cleaned and binned similarly to the association analysis for the classification model to improve model accuracy. Binning segregates the data to reduce its spread and thus helps in classifying it better.

Parameters –

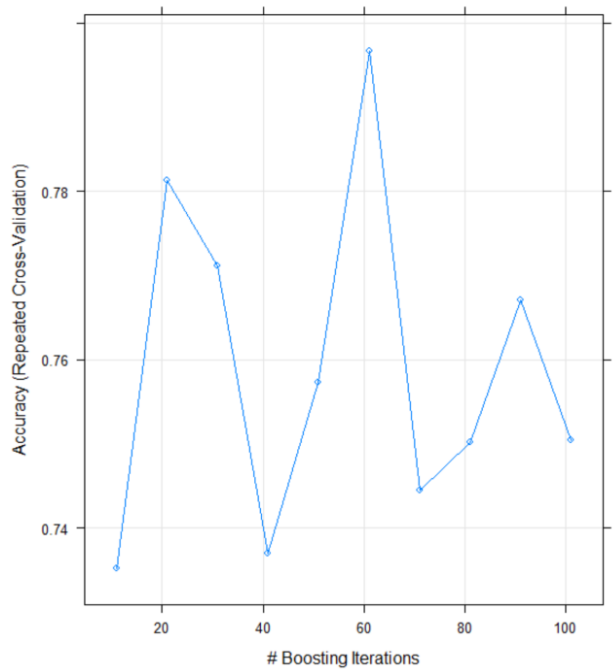
- Cross-validation parameters – 10-fold, 5 times
- Train data – 75% (random)
- Test data – 25% (remaining)
- Algorithm – LogitBoost
- tuneLength = 10

```
analysisColumns = c("Airline.Status", "Type.of.Travel",
"Eating.and.Drinking.at.Airport", "Departure.Delay.in.Minutes",
"Flights.Per.Year", "Price.Sensitivity", "olong", "dlat",
"Total.Freq.Flyer.Accts")
analysisData = dfBinnedData[, names(dfBinnedData) %in% c(analysisColumns,
"Likelihood.to.recommend") ]
analysisData$Likelihood.to.recommend[is.na(analysisData$Likelihood.to.recomme
nd)] = "Passive"

analysisData = prepareForAnalysis(analysisData, analysisColumns)

fitControl <- trainControl(method = "repeatedcv", number = 10, repeats = 5)

set.seed(1)
inTraining = createDataPartition(analysisData$Likelihood.to.recommend, p =
0.75, list = FALSE)
trainData = analysisData[inTraining,]
testData = analysisData[-inTraining,]

logitBoost <- train(factor(Likelihood.to.recommend) ~., data = trainData,
method = "LogitBoost", trControl=fitControl, preProcess = c("center",
"scale"), tuneLength = 10)
```

Results –

The train accuracy was about 79.55%

```
plot(logitBoost)
```



The test accuracy was about 80.48%.

```
result = predict(logitBoost, testData)
sum(result ==
testData$Likelihood.to.recommend)/length(testData$Likelihood.to.recommend)
```

The F1 score was about 0.87.

```
F1_Score(testData$Likelihood.to.recommend, result)
```

```
> varImp(logitBoost)
ROC curve variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 34)

                                        Detractor Passive Promoter
Type.of.Travel_Personal Travel           100.0000  75.255  100.000
Airline.Status_Silver                     39.1323  26.209   39.132
Flights.Per.Year_>40                      22.3341  17.303   22.334
Price.Sensitivity_1                       20.4935  13.811   20.494
Price.Sensitivity_2                       19.3876  12.175   19.388
Total.Freq.Flyer.Accts_ 2                 13.0589   7.584   13.059
Airline.Status_Gold                       11.3982   7.838   11.398
olong_>-95                                10.3662  10.117   10.366
Total.Freq.Flyer.Accts_ 1                 10.1912  10.191    8.929
Type.of.Travel_Mileage tickets             6.8661   6.866    4.093
Airline.Status_Platinum                    4.1670   6.149    6.149
Eating.and.Drinking.at.Airport_200-300     3.1111   6.078    6.078
Eating.and.Drinking.at.Airport_100-200     5.0762   5.829    5.829
dlat_30-45                                 3.1780   3.178    3.074
olong_-145 - -120                          3.0570   2.008    3.057
Total.Freq.Flyer.Accts_ 3                  2.4282   1.652    2.428
Price.Sensitivity_3                        2.2865   2.398    2.398
Departure.Delay.in.Minutes_100-200         2.2134   1.147    2.213
dlat_45-60                                 0.7239   1.189    1.189
Eating.and.Drinking.at.Airport_300-400     0.8902   1.172    1.172
```

The strongest predictor was the dummy variable Type.of.Travel_Personal Travel.

The prediction of personal travel passengers was detractors for about 85.4% of the time.

```
> personalTravelResult = predict(logitBoost, testData[testData$`Type.of.Travel_Personal Travel` == 1,])
> sum(personalTravelResult == "Detractor")/nrow(testData[testData$`Type.of.Travel_Personal Travel` == 1,])
[1] 0.854321
```

Conclusion – Passengers with travel type as personal, for a large proportion in the dataset, are and according to the classification model (high accuracy, predicted value), predicted to be detractors and the airline should focus more on them. Further analysis was done on this subset of data to gain more insight into which features strongly affect their ratings.

Personal Travel Data Analysis

The data with the value for Type.of.Travel as "Personal Travel" was extracted (3212 rows) and a combination of association rules mining and classification modeling was used to gain more insight for these type of passengers.

Data munging and feature addition

The NA values and the columns not needed were removed in a similar manner as described in the exploratory analysis section. The numerical columns were also binned in the same way with the Likelihood.to.recommend discretized into *detractor, passive* and *promoter*.

Association rules mining (all personal travel data)

2 rulesets were considered –

rhs = "Likelihood.to.recommend=Promoter", support = 0.005, confidence = 0.5 (24 rules)

rhs = "Likelihood.to.recommend=Detractor", support = 0.05, confidence = 0.8, lift = 1.178 (17 rules)

```
rulesetPromoters <- apriori(personalTravelTnx,
                    parameter=list(support=0.005,confidence=0.5),
                    appearance = list(default="lhs", rhs=("Likelihood.to.rec-
ommend=Promoter")))

rulesetDetractors <- apriori(personalTravelTnx,
                             parameter=list(support=0.05,confidence=0.8),
                             appearance = list(default="lhs", rhs=("Likeli-
hood.to.recommend=Detractor")))

arules::inspect(rulesetPromoters)
arules::inspect(rulesetDetractors[quality(rulesetDetractors)$lift > 1.178])
```

Promoters (top 6 rules) –

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[1]	{Airline.Status=Silver,olong=>-95,olat=<30}	{Likelihood.to.recommend=Promoter}	0.006	0.529	8.336	18.000
[2]	{Age=35-45,Gender=Male,Price.Sensitivity=1,Flights.Per.Year=<40}	{Likelihood.to.recommend=Promoter}	0.006	0.513	8.074	20.000
[3]	{Airline.Status=Silver,Flights.Per.Year=<40,olong=>-95,olat=<30}	{Likelihood.to.recommend=Promoter}	0.006	0.529	8.336	18.000
[4]	{Airline.Status=Silver,Shopping.Amount.at.Airport=<100,olong=>-95,olat=<30}	{Likelihood.to.recommend=Promoter}	0.005	0.567	8.922	17.000
[5]	{Airline.Status=Silver,Flight.cancelled=No,olong=>-95,olat=<30}	{Likelihood.to.recommend=Promoter}	0.006	0.529	8.336	18.000
[6]	{Airline.Status=Silver,Departure.Delay.in.Minutes=<100,olong=>-95,olat=<30}	{Likelihood.to.recommend=Promoter}	0.005	0.531	8.365	17.000

Detractors (top 6 rules) –

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	,	All
[1]	{Airline.Status=Blue,Origin.State=Texas,olat=<30}	{Likelihood.to.recommend=Detractor}	0.050	0.994	1.184	161.000
[2]	{Airline.Status=Blue,Eating.and.Drinking.at.Airport=<100,Origin.State=Texas}	{Likelihood.to.recommend=Detractor}	0.058	0.989	1.178	186.000
[3]	{Airline.Status=Blue,olong=-120 - -95,olat=<30}	{Likelihood.to.recommend=Detractor}	0.050	0.994	1.184	161.000
[4]	{Airline.Status=Blue,Origin.State=Texas,olong=-120 - -95,olat=<30}	{Likelihood.to.recommend=Detractor}	0.050	0.994	1.184	161.000
[5]	{Airline.Status=Blue,Eating.and.Drinking.at.Airport=<100,Origin.State=Texas,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.058	0.989	1.178	186.000
[6]	{Airline.Status=Blue,Loyalty=<-0.5,Eating.and.Drinking.at.Airport=<100,Class=Eco,Day.of.Month=before 15th,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.058	0.989	1.178	187.000

Inferences –

- There seemed to a weak association of gender and airline status.
- Male passengers on airlines with status as “Silver” tend to be promoters.
- Passengers on airlines with status as “Blue” tend to be detractors.

No definite conclusions could be made from this analysis as the results were similar to what was previously found. A subset of this data where the Airline.Status was “Blue” was considered for further analysis. This subset contained 2504 rows and 89.85% of the passengers were detractors.

Association rules mining (personal travel, blue airline data)

2 rulesets were considered –

rhs = “Likelihood.to.recommend=Promoter”, support = 0.002, confidence = 0.5, lift = 28 (12 rules)

rhs = “Likelihood.to.recommend=Detractor”, support = 0.05, confidence = 0.8, lift = 1.11 (20 rules)

```
rulesetPromoters <- apriori(personalBlueTnx,
                            parameter=list(support=0.002,confidence=0.5),
                            appearance = list(default="lhs", rhs=("Likeli-
hood.to.recommend=Promoter")))

rulesetDetractors <- apriori(personalBlueTnx,
                             parameter=list(support=0.05,confidence=0.8),
                             appearance = list(default="lhs", rhs=("Likeli-
hood.to.recommend=Detractor")))

arules::inspect(rulesetPromoters[quality(rulesetPromoters)$lift > 28])
arules::inspect(rulesetDetractors[quality(rulesetDetractors)$lift > 1.11])
```

Promoters (top 6 rules) –

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	A	All
[1]	{Age=35-45,Gender=Male,Price.Sensitivity=1,Total.Freq.Flyer.Accts=2,Day.of.Month=after 15th}	{Likelihood.to.recommend=Promoter}	0.002	0.750	31.300	6.000
[2]	{Age=35-45,Gender=Male,Total.Freq.Flyer.Accts=2,Day.of.Month=after 15th,Flight.Distance=<1000}	{Likelihood.to.recommend=Promoter}	0.002	0.750	31.300	6.000
[3]	{Age=35-45,Gender=Male,Total.Freq.Flyer.Accts=2,Eating.and.Drinking.at.Airport=<100,Day.of.Month=after 15th}	{Likelihood.to.recommend=Promoter}	0.002	0.750	31.300	6.000
[4]	{Age=35-45,Gender=Male,Total.Freq.Flyer.Accts=2,Day.of.Month=after 15th,dlat=30-45}	{Likelihood.to.recommend=Promoter}	0.002	0.750	31.300	6.000
[5]	{Age=35-45,Gender=Male,Total.Freq.Flyer.Accts=2,Day.of.Month=after 15th,Flight.cancelled=No}	{Likelihood.to.recommend=Promoter}	0.002	0.750	31.300	6.000
[6]	{Age=35-45,Gender=Male,Price.Sensitivity=1,Total.Freq.Flyer.Accts=2,dlat=30-45}	{Likelihood.to.recommend=Promoter}	0.002	0.750	31.300	6.000

Detractors (top 6 rules) –

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[1]	{Year.of.First.Flight=after 2009,Loyalty=<-0.5,Shopping.Amount.at.Airport=<100,Eating.and.Drinking.at.Airport=<100,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.050	1.000	1.113	126.000
[2]	{Gender=Female,Total.Freq.Flyer.Accts=0,Day.of.Month=before 15th,Flight.Distance=<1000,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.052	1.000	1.113	130.000
[3]	{Gender=Female,Total.Freq.Flyer.Accts=0,Class=Eco,Day.of.Month=before 15th,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.056	1.000	1.113	139.000
[4]	{Gender=Female,Total.Freq.Flyer.Accts=0,Day.of.Month=before 15th,Flight.cancelled=No,Flight.Distance=<1000,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.051	1.000	1.113	128.000
[5]	{Gender=Female,Total.Freq.Flyer.Accts=0,Day.of.Month=before 15th,Arrival.Delay.in.Minutes=<100,Flight.Distance=<1000,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.050	1.000	1.113	126.000
[6]	{Gender=Female,Total.Freq.Flyer.Accts=0,Day.of.Month=before 15th,Departure.Delay.in.Minutes=<100,Flight.Distance=<1000,olong=-120 - -95}	{Likelihood.to.recommend=Detractor}	0.051	1.000	1.113	127.000

Inferences –

- There seemed to be a strong association of age and gender for promoters, and total frequent flyer accounts and gender for detractors.
- Male passengers between the age of 30-45 tend to be promoters.
- Female passengers with no frequent flyer accounts tend to be detractors.

The columns with high association with each other were considered for classification modeling –

Age

Gender

Flight.Distance

Eating.and.Drinking.at.Airport

olong

Arrival.Delay.in.Minutes

Loyalty

Total.Freq.Flyer.Accts

Classification model (personal travel, blue airline data)

The columns were cleaned and binned similarly to the association analysis for the classification model to improve model accuracy.

Parameters –

- Cross-validation parameters – 10-fold, 5 times
- Train data – 75% (random)

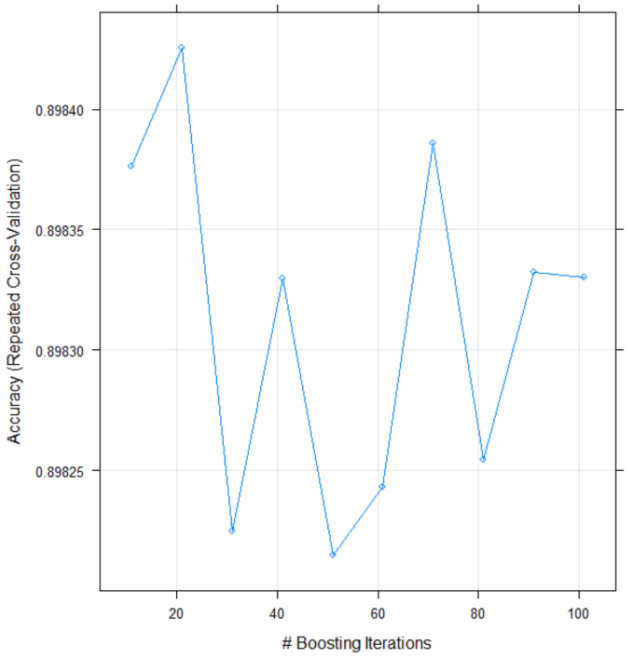
- Test data – 25% (remaining)
- Algorithm – LogitBoost, svmRadial
- tuneLength = 10

```
analysisColumns1 = c("Age", "Gender", "Flight.Distance", "Eating.and.Drink-
ing.at.Airport",
                     "olong", "Arrival.Delay.in.Minutes", "Loyalty", "To-
tal.Freq.Flyer.Accts")
analysisData1 = personalBlueBinned[, names(personalBlueBinned) %in% c(analy-
sisColumns1, "Likelihood.to.recommend")]
analysisData1 = prepareForAnalysis(analysisData1, analysisColumns1)

set.seed(1)
inTraining1 = createDataPartition(analysisData1$Likelihood.to.recommend, p =
.75, list = FALSE)
trainData1 = analysisData1[inTraining1,]
testData1 = analysisData1[-inTraining1,]

fitControl <- trainControl(method = "repeatedcv", number = 10, repeats = 5)
```

LogitBoost –

The train accuracy was around 89.85%.

```
logitBoost1 <- train(factor(Likelihood.to.recommend) ~., data = trainData1,
method = "LogitBoost", trControl=fitControl, preProcess = c("center",
"scale"), tuneLength = 10)
plot(logitBoost1)
```



The test accuracy was about 89.76%.

```
result5 = predict(logitBoost1, newdata = testData1)
sum(result5 ==
testData1$Likelihood.to.recommend)/length(testData1$Likelihood.to.recommend)
```

Both the above calculated train and test accuracies are not a good metric for measuring the accuracy of the model as the data is heavily skewed with respect to the classes (around 89% are detractors).

The F1 score was about 0.95. It provides a better metric for measuring the accuracy of the model as it uses a combination of precision and recall to evaluate the accuracy.

```
F1_Score(testData1$Likelihood.to.recommend, result5)
```

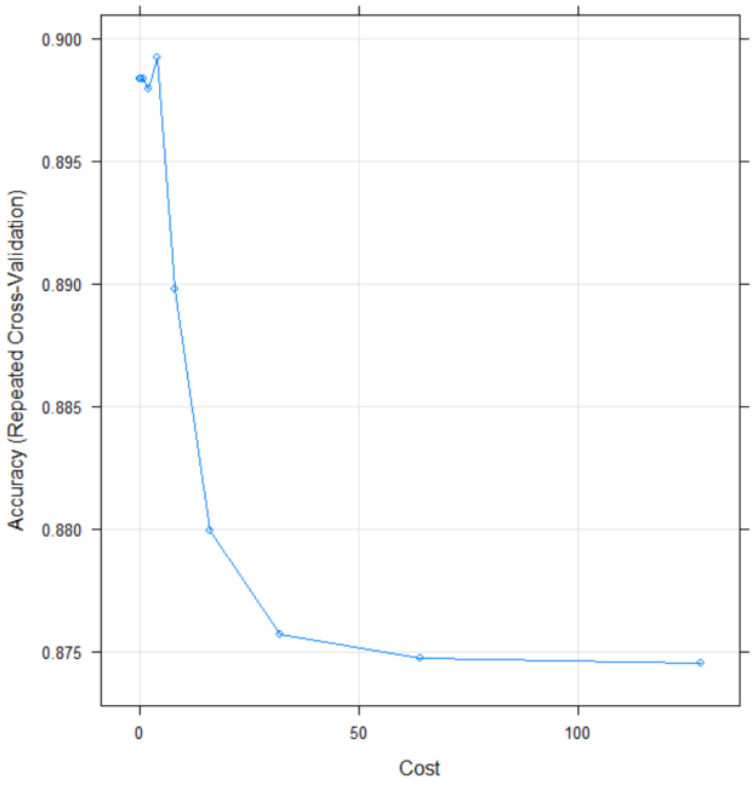
```
> varImp(logitBoost1)
ROC curve variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 31)

                                      Detractor Passive Promoter
Gender_Male                             100.000  96.517  100.000
Loyalty_<-0.5                            91.614  55.944   91.614
Age_35-45                                79.225  70.014   79.225
Total.Freq.Flyer.Accts_2                 77.616  48.259   77.616
Age_65-75                                71.455  73.934   73.934
Age_>75                                  71.350  54.664   71.350
Age_25-35                                68.800  54.498   68.800
Age_45-55                                35.374  42.737   42.737
Eating.and.Drinking.at.Airport_100-200   40.006  40.006   26.228
olong_-145 - -120                        36.811  39.811   39.811
Flight.Distance_1000-2000                32.977  36.719   36.719
Loyalty_0-0.5                            35.264  24.737   35.264
Age_55-65                                33.516  18.553   33.516
Total.Freq.Flyer.Accts_3                 24.404  27.884   27.884
Loyalty_>0.5                             25.694  20.430   25.694
Total.Freq.Flyer.Accts_1                 23.554  15.405   23.554
Flight.Distance_>2000                    21.290  23.522   23.522
olong_>-95                               16.290   9.830   16.290
Arrival.Delay.in.Minutes_100-200          9.671   7.186    9.671
Eating.and.Drinking.at.Airport_200-300    5.803   4.969    5.803
```

The strongest predictor was the dummy variable for Gender.

svmRadial –

The train accuracy was about 89.92%.

```
svmRadial1 <- train(factor(Likelihood.to.recommend) ~., data = trainData1,
method = "svmRadial", trControl=fitControl, preProcess = c("center",
"scale"), tuneLength = 10)
plot(svmRadial1)
```



The test accuracy was about 89.44%.

```
Result6 = predict(svmRadial1, newdata = testData1)
sum(result6 ==
testData1$Likelihood.to.recommend)/length(testData1$Likelihood.to.recommend)
```

The F1 score was about 0.94.

```
F1_Score(testData1$Likelihood.to.recommend, result6)
```

```
> varImp(svmRadial1)
ROC curve variable importance

  variables are sorted by maximum importance across the classes
  only 20 most important variables shown (out of 31)

                                        Detractor Passive Promoter
Gender_Male                               100.000  96.517  100.000
Loyalty_<-0.5                              91.614  55.944   91.614
Age_35-45                                  79.225  70.014   79.225
Total.Freq.Flyer.Accts_2                   77.616  48.259   77.616
Age_65-75                                  71.455  73.934   73.934
Age_>75                                    71.350  54.664   71.350
Age_25-35                                  68.800  54.498   68.800
Age_45-55                                  35.374  42.737   42.737
Eating.and.Drinking.at.Airport_100-200     40.006  40.006   26.228
olong_-145 - -120                          36.811  39.811   39.811
Flight.Distance_1000-2000                  32.977  36.719   36.719
Loyalty_0-0.5                              35.264  24.737   35.264
Age_55-65                                  33.516  18.553   33.516
Total.Freq.Flyer.Accts_3                   24.404  27.884   27.884
Loyalty_>0.5                               25.694  20.430   25.694
Total.Freq.Flyer.Accts_1                   23.554  15.405   23.554
Flight.Distance_>2000                      21.290  23.522   23.522
olong_>-95                                 16.290   9.830   16.290
Arrival.Delay.in.Minutes_100-200            9.671   7.186    9.671
Eating.and.Drinking.at.Airport_200-300      5.803   4.969    5.803
```

The strongest predictor was the dummy variable for Gender.

The prediction of female passengers was detractors for about 99.25% of the time.

```
> femalePredict = predict(svmRadial1, testData1[testData1$Gender_Male == 0,])
> sum(femalePredict == "Detractor")/nrow(testData1[testData1$Gender_Male == 0,])
[1] 0.9925
```

Conclusion – Female passengers with travel type as personal and on airlines with status as Blue, for a large proportion in the dataset, are and according to the classification model (high accuracy, predicted value), are predicted to be detractors and the airline should focus more on them. No definite conclusions could be made about the factors which negatively affect their ratings. The features which show a strong dependency on the Likelihood.to.recommend column do not tell us more about the passengers' experience or the quality of the airlines.

Overall Analysis Conclusions (Personal Travel Data)

Passenger Category	Proportion	Average rating
Personal travel	31.24%	5.28 (vs 7.17 overall)
Personal travel, Blue airline	24.35%	4.86 (vs 7.17 overall)
Personal travel, Blue airline, Female	15.70%	4.78 (vs 7.17 overall & 6.98 overall female)

The above table summarizes the distribution of the personal travel passengers. There are 3 major take backs from the data and the above analyses – firstly, the passengers with travel type as personal tend to be detractors for a large proportion of the time; secondly, the passengers using airlines with Blue status in this category tend to be detractors for most of the time, and lastly, the female passengers of this sub-category tend to be detractors almost all the time. No definite conclusions could be made about how the airline can improve the ratings of these passengers, but it would benefit heavily from focusing on these passengers as their contributions to the overall ratings of the airline is significant as can be seen from the table above by comparing the number of these passengers and their average ratings to the overall values.

Sentiment analysis on customer reviews

We have 265 comments included 7500 words. And words cloud according to the frequency is below.

We find word cloud includes all words can't help us to analyze what the customers' concentration. So, we analyze the positive words and negative words.

There are 2091positive words and 2091 negative words. From positive words cloud, our customers think our comfortable and friendly service. From negative words, our customers feel unsatisfied with flight delay.

The words cloud after deleting the stop words and sentiment words to find the words most mentioned about service, seats, time, food, luggage from our customers.

ACTIONABLE INSIGHTS

Based on all three models, it is clear that personal travelers with airline status Blue are the major detractors. Southeast Airlines can concentrate on these passengers in order to increase their NPS.

The recommendations based on our analysis are-

1. Improve quality of service for passengers with airline status Blue and price sensitivity 1 by providing promotional offers or discounts.
2. Customers above the age of 60 years and flight delays tend to give bad ratings. Comfort level for them can be increased in the airport by establishing essential infrastructure.
3. Inflight services can be enhanced for personal travelers.
4. Shopping options and discounts at the airport for female customers traveling with top three airline partners can increase likelihood to recommend.
5. From sentiment analysis it is evident that customers are concerned about time, food and luggage. Providing more food options inflight and free check-in luggage will increase number of promoters.