Chiyoung Lii, Dahai Alex Liu, Gao Ge

Clii01@syr.edu, dliu24@syr.edu, ggao05@syr.edu

Abstract

IST 687 Final Project Report, Due on 4/25/2019

Predicting airline customer satisfaction

Analysis and Synthesis of Airline Customer Data

Contents

[Introduction 2](#_Toc6846970)

[Data Transformations 3](#_Toc6846971)

[Descriptive Statistics and Diagnostic Visualizations 4](#_Toc6846972)

[Pearson Correlation, and finding Interesting Variables 4](#_Toc6846973)

[Satisfaction Data Summary 5](#_Toc6846974)

[Flights Per Year 6](#_Toc6846975)

[Loyalty 7](#_Toc6846976)

[Price Sensitivity Graphs 8](#_Toc6846977)

[Arrival Delay in Minutes 9](#_Toc6846978)

[Airline Status, Types of Travel, Class of Flight 11](#_Toc6846979)

[Model Development and Results 13](#_Toc6846980)

[Support Vector Machine Modeling 13](#_Toc6846981)

[Confusion Matrixes 13](#_Toc6846982)

[SVM Models without bucketing 14](#_Toc6846983)

[Analysis results 14](#_Toc6846984)

[Association Rule Mining 15](#_Toc6846985)

[Linear modeling 18](#_Toc6846986)

[Analysis results 19](#_Toc6846987)

[Actionable Insights and Conclusion 20](#_Toc6846988)

[Appendix 21](#_Toc6846989)

[Airline Status, Types of Travel, Class of Flight Code 21](#_Toc6846990)

[Pearson Correlation Coefficient Code 22](#_Toc6846991)

[SVM Modeling Code (In R) 23](#_Toc6846992)

[Linear Model 2 Code 26](#_Toc6846993)

[Satisfaction Data Discovery Code 28](#_Toc6846994)

[Association Rule Code 29](#_Toc6846995)

[SVM Code without Bucketing 31](#_Toc6846996)

[Bibliography 33](#_Toc6846997)

# Introduction

The customer satisfaction is critically important for customer retention (Anders Gustafsson, 2005). Because commercial airplanes are generally the same model, customer satisfaction is even more important in the airline industry as a method of customer retention. This report investigates airline customer survey data to discover what important factors determine customer satisfaction and dissatisfaction. The processed and analyzed data will be adapted into visual form to better convey the actionable insights. Finally, research is done to produce actionable insights that airlines can use to increase customer satisfaction and reduce customer dissatisfaction.

Some data questions we wanted to answer were:

1. What factors contribute to Customer Satisfaction? How can they be changed to increased customer satisfaction?
2. What factors contribute to Customer Dissatisfaction? How can they be changed to reduce customer dissatisfaction?
3. Which factors are more important than others? Rank the factors by level of impact. Is there a general trend that can be gleaned from the data?
4. What is the 0.5 and 0.95 percentile of the satisfaction levels? How is this important when determining and evaluating what actions to take?
5. How do the factors interact? What happens when multiple factors are prevalent in a single experience?

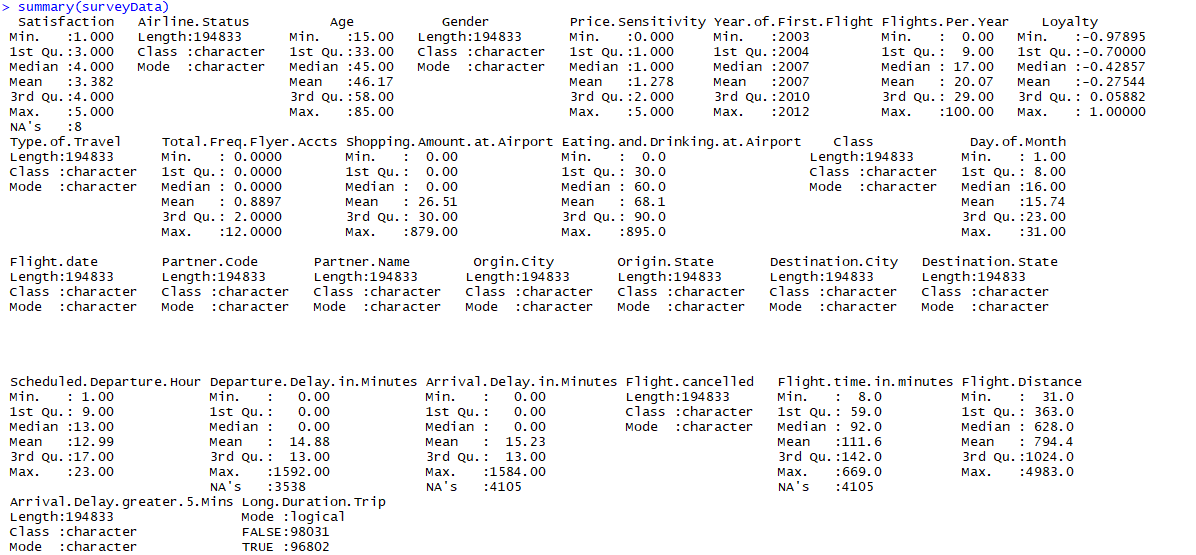
This report will talk about how the data was acquired, and comment on the quality of the data. It will also talk about our project management structure, and finally comment on candidate variables for analysis and why they are candidate variables.

The primary purpose of this project update report is to predict low customer satisfaction using regression and classification methods, learning from the dataset provided, in order to generate actionable business insights to increase customer satisfactions. Also marketing strategies can be generated based on the finding of correlations between variables and customer satisfaction.

# Data Discovery and Exploration

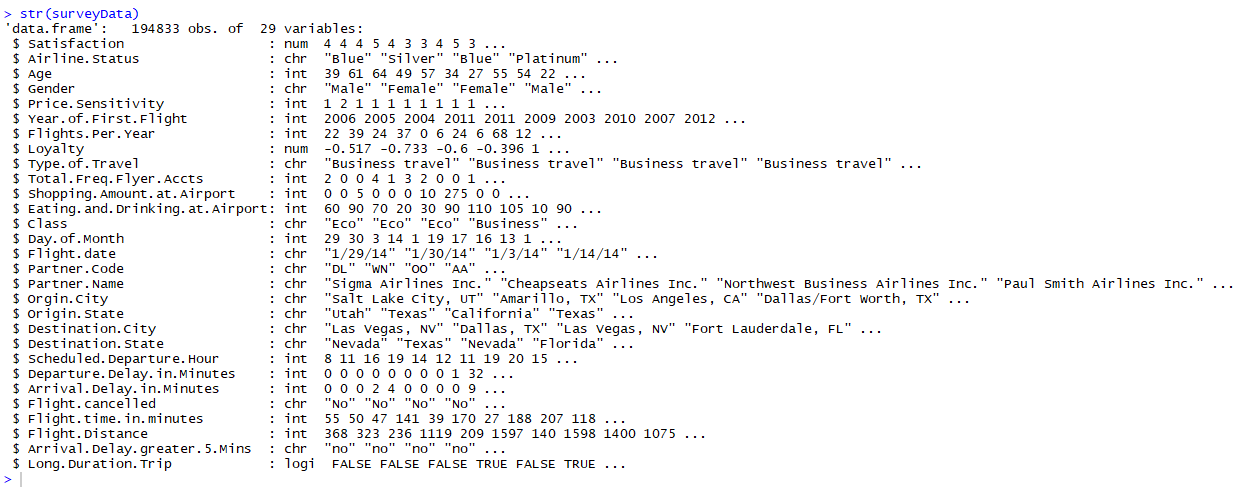
The data was passed through at least three different entities before becoming available to us. Firstly, the data was written by customers and given to the airlines. The data was then collated by some entity, and finally that entity passed the data on to Syracuse University, where it was then passed to us. Because the data acquisition chain is so long, we predict some problems with the data itself.

Image 1: Screenshot of the summary() of the survey data in RStudio. Note the NAs



A simple look at a summary of the data using RStudio’s summary() command indicates a few issues with data integrity. For example, in the Satisfaction column there are 8 NAs, and the Arrival.Delay.In.Minutes column there are 4105 NAs. We removed all the invalid data to make sure our models are correct.

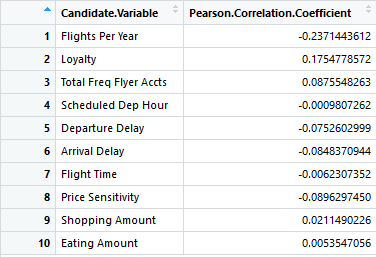
A look at the structure of the data using str() shows another couple of issues. For example, the Origin.City is a multivalued datatype, with the State put in after the city. This is also redundant because the Origin.State is in a column right below the Origin.City. Same problem with Destination.City.

Image 2: Screenshot of str() of the survey data in RStudio. The City data will have to be cleaned.

# Descriptive Statistics and Diagnostic Visualizations

## Pearson Correlation, and finding Interesting Variables

After preliminary candidate variable testing showed some variables were unsuitable (such as gender or age, things which cannot be affected, we calculated the Pearson Correlation Coefficient for some other variables of interest and the Satisfaction variable. This code is in the appendix, section “Pearson Correlation Coefficient Code”. The results are displayed below:



A table calculated Pearson Correlation Coefficient for selected candidate variables vs a Satisfaction Variable.

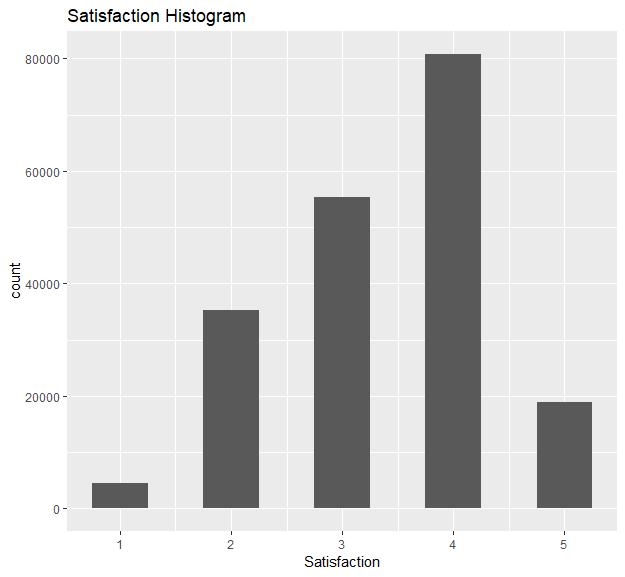
Ranking by absolute value of correlation:

1. **Flights Per Year**
2. **Loyalty**
3. **Price Sensitivity**
4. **Total Frequent Flyer**
5. **Arrival Delay**
6. Departure Delay
7. Shopping Amount
8. Flight Time
9. Eating Amount
10. Scheduled Departure Time

The results showed that Flights Per Year and Loyalty had the highest absolute possible impact on determining Satisfaction, with Flights Per Year having a negative impact and Loyalty having a positive one. Next, Total Freq Flyer Accts and Arrival Delay also had a significant possibility of having a positive impact on the satisfaction of customers. Some variables were discovered to be unimpactful (such as shopping and eating amount spent at airport). Due to these results, we will be including these variables in our data models.

Though we did not calculate the Pearson Correlation for Airline Status, Types of Travel, and Class of Flight, we still recognized them as important factors and did a separate analysis for them as well.

## Satisfaction Data Summary

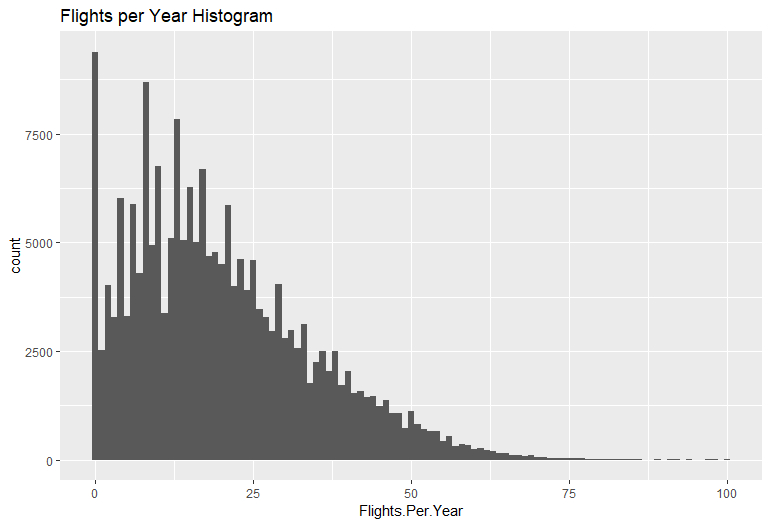
 There were 8 NA’s in the satisfaction column. These were not considered for our data exploration. The code use to determine the percentiles, quartiles, median and mean are in the appendix section “Satisfaction Data Discovery Code”.

The data seems to indicate that most people were relatively satisfied with their experience, as there seems to be a normal curve centered around 4 satisfaction.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 5th percentile | 1st Quartile | Median | Mean | 3rd Quartile | 95th percentile |
| 2 | 3.000 | 4.000 | 3.382 | 4.000 | 5 |

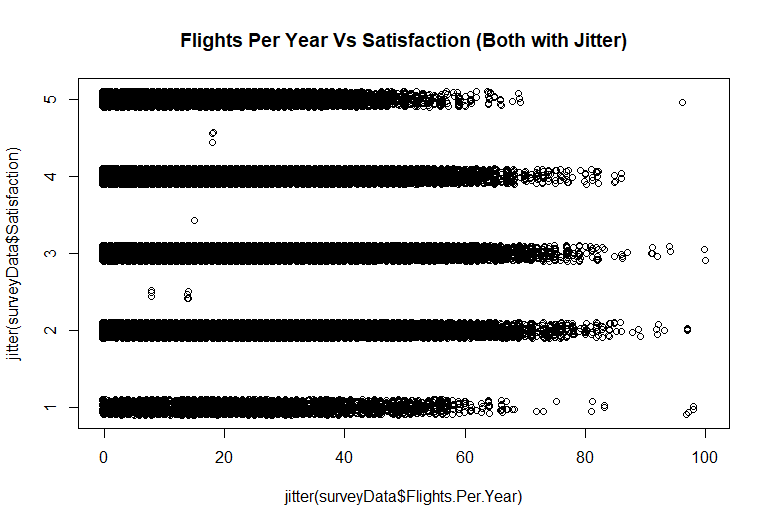
The 5th and 95th percentile also indicate that very few people gave the worst score or the best score. However, there is a strong skew to the higher end of the graph, which implies a bias towards positive reviews.

## Flights Per Year

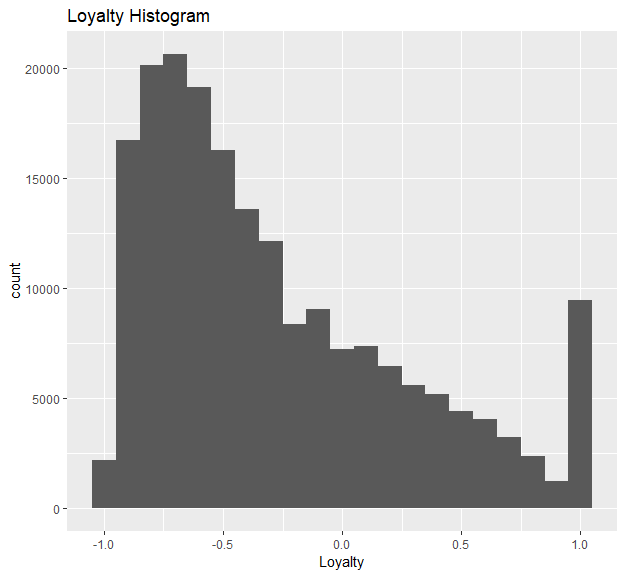
The histogram for Flights per Year revealed a strong normal curve with the median flights per year at 17 flights.

The below scatter plot of Flights per Year vs Satisfaction seems to indicate that most people gave a three or a two of satisfaction, with a bit of a skew towards two. It also seems that a few people with more than a hundred flights gave 1 satisfaction, which is very rare.

|  |  |  |  |
| --- | --- | --- | --- |
| 1st Quartile | Median | Mean | 3rd Quartile |
| 9 | 17 | 20 | 29 |



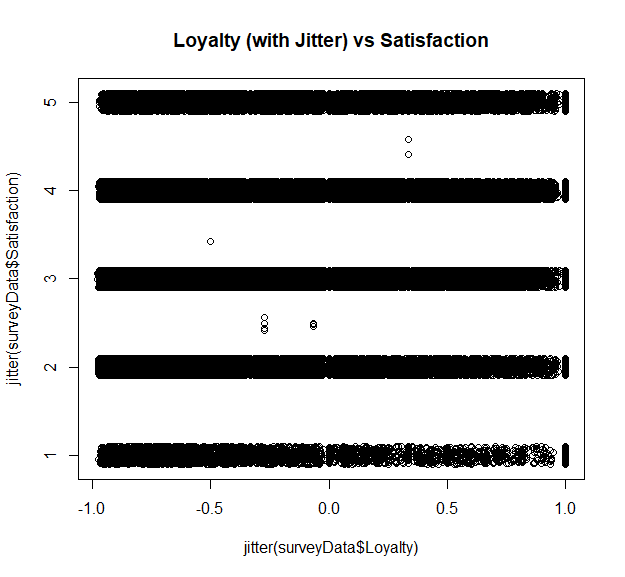
## Loyalty



The histogram looked almost like a j-shaped curve, however, it is important to note that the first area (around x = -0.75) of the curve is much higher than the then the end part of the curve (around x = 1.0) It doesn’t seem like many people are loyal to airline companies.

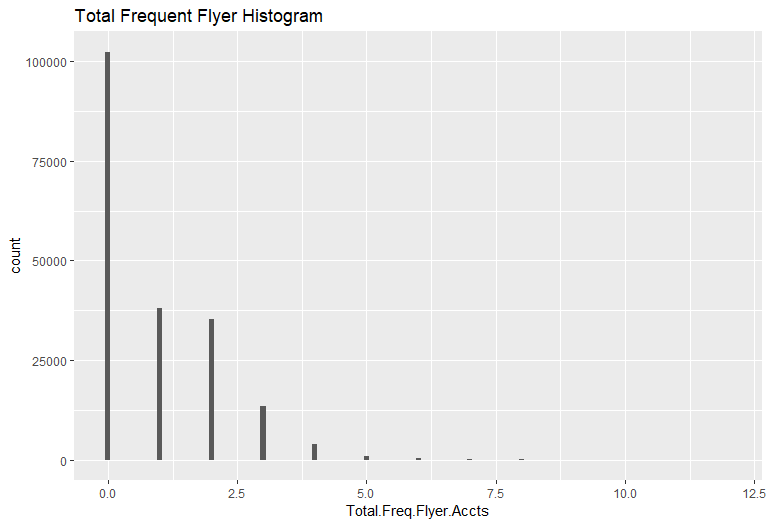
It makes sense to think that if people are more loyal to the airport, they would be biased towards being satisfied with the data. Modeling the loyalty predicting the customer satisfaction with a simple linear model would probably be the best bet for handling this data.

|  |  |  |  |
| --- | --- | --- | --- |
| 1st Quartile | Median | Mean | 3rd Quartile |
| -0.7 | -0.42857 | -0.27544 | 0.05882 |



This data would probably best be modeled with a linear or stepwise model, due to the nice and clean continuous numeric values.

## Total Frequent Flyer Accounts

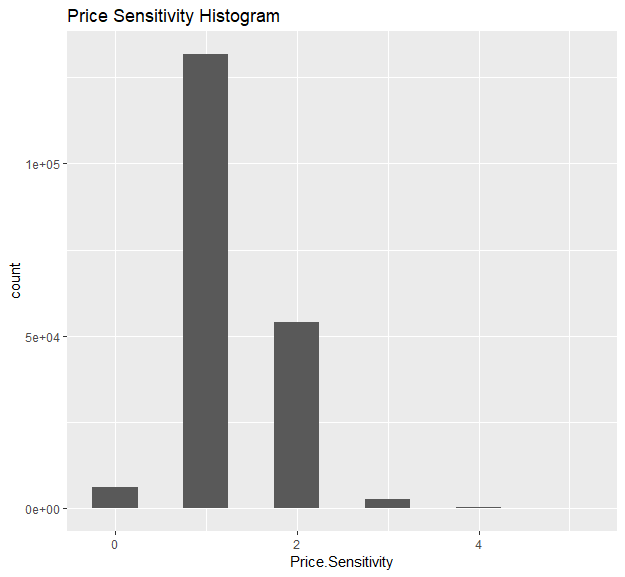
This looks like a clear case of linear or exponential decay. While many customers have at least 1 frequent flyer account, most customers have 0 total frequent flyer accounts, and very, very few have more than three.

The scatter plot of total frequent flyer accounts vs satisfaction seems to indicate there is a central tendency located at 1 account and 3 satisfaction.

|  |  |  |  |
| --- | --- | --- | --- |
| 1st Quartile | Median | Mean | 3rd Quartile |
| 0.0 | 0.0 | 0.8897 | 2.0 |

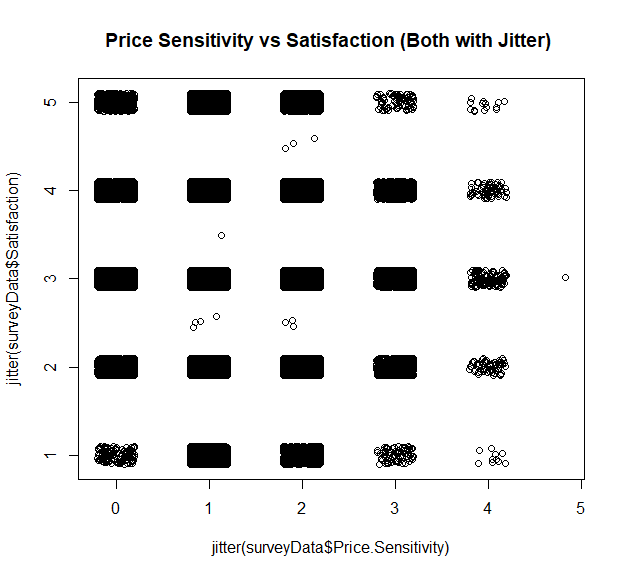


## Price Sensitivity Graphs

A histogram was made to understand the general trend of the data. The data was roughly a normal distribution with a mean around 1. This makes sense because few people would be willing to say they are really sensitive to price, for fear of looking like a miser.

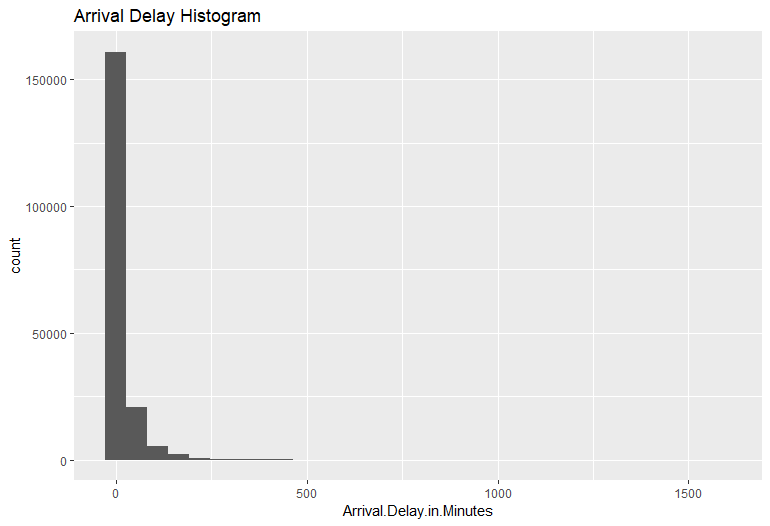
However, it is reasonable to suspect that people who are willing to be more sensitive to price, would be more willing to express dissatisfaction with their airline. It would be worthwhile to transform this data into two different classes (low price sensitivity and high sensitivity) and attempting to predict customer satisfaction that way.

|  |  |  |  |
| --- | --- | --- | --- |
| 1st Quartile | Median | Mean | 3rd Quartile |
| 1.000 | 1.000 | 1.278 | 2.000 |



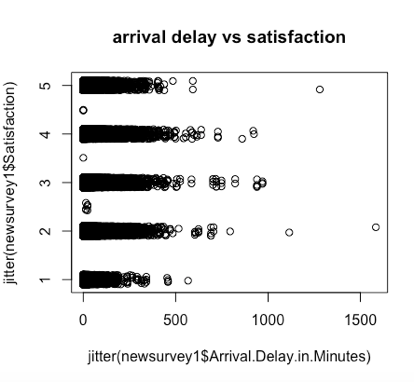
A short bar plot was created. The data seems fairly consistent when the price sensitivity is 0, 1 or 2. It seems like when people had a price sensitivity of 3 or 4, they were less likely to give a satisfaction of 1 or 5, probably due to there being less data points at 1 and 5. The plot may indicate a central tendency at (1, 3).

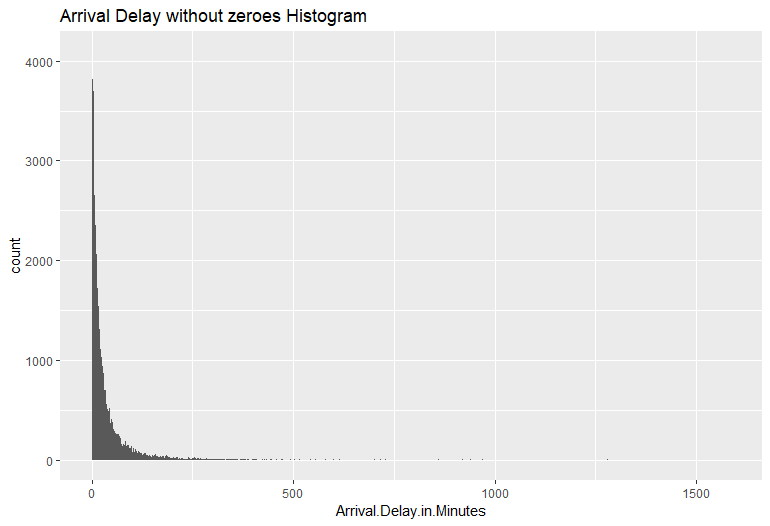
## Arrival Delay in Minutes

This histogram reveals the vast majority of the reported Arrival Delay was zero. Because of this, the 1st quartile, median and mean are all zero. To focus on the important data, all zeroes were stripped from the vector and the histogram was recreated.

|  |  |  |  |
| --- | --- | --- | --- |
| 1st Quartile | Median | Mean | 3rd Quartile |
| 0 | 0 | 0 | 0 |

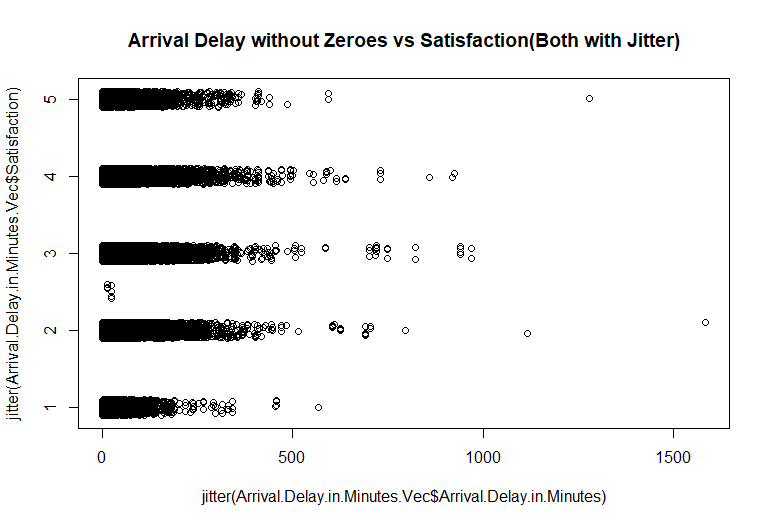
A scatter plot of arrival delay in minutes vs customer satisfaction was also made. From the picture we can see that when the arrival delay is less than 500 minutes, there is no obvious trend about satisfaction, which means the distribution of satisfaction is evenly. But when the arrival delay is more than 500 minutes, people is more likely to choose the medium level of satisfaction such as 2, 3 and4.



The arrival delay data without any zeroes still showed a strong asymptotic behavior at zero, but the central tendencies are much more interesting. This seems to indicate that if there was a delay, it usually lasted no longer than 17 minutes. It is possible that this would result in a strong loss of customer satisfaction.

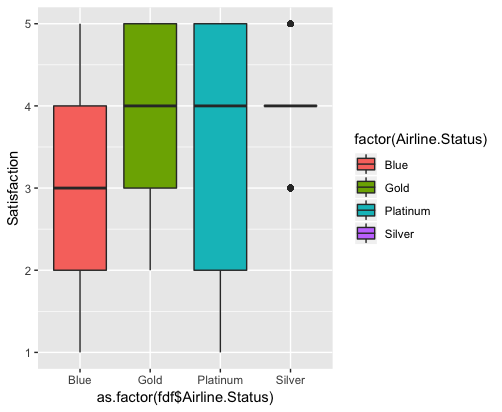
|  |  |  |  |
| --- | --- | --- | --- |
| 1st Quartile | Median | Mean | 3rd Quartile |
| 6.00 | 17.00 | 34.18 | 40.00 |

Not much difference was seen in the scatter plot when the zeroes were taken out.

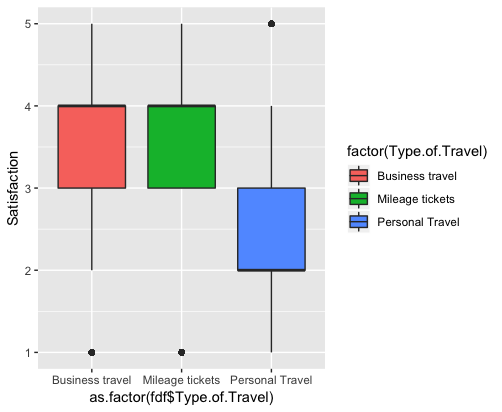


## Airline Status, Types of Travel, Class of Flight

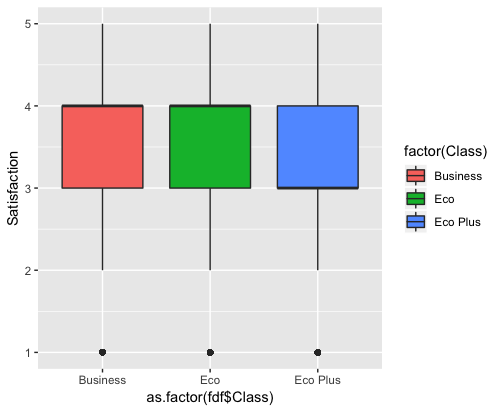
We used a boxplot rather than a histogram to compare the distribution of satisfaction level and compare the median and mean. The code is in the appendix under “Airline Status, Types of Travel, Class of Flight Code:



Based on this output, we can see that the average satisfaction level increases with higher membership status. For the Airline Status variable, the summary indicates that silver members have the highest satisfaction and blue members have the lowest satisfaction. Gold and platinum in the middle. Thus, company should increase the number of silver member to increase their satisfaction.



Business and mileage travel generally produce a higher level of satisfaction than regular personal travels.



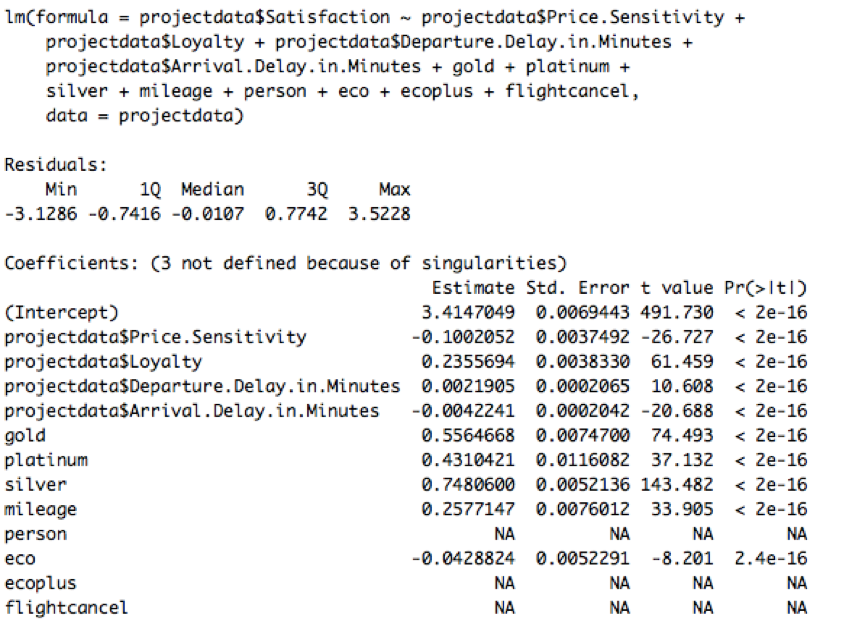
For flight ticket class, business class has the highest satisfaction, and surprisingly the eco-plus is the lowest satisfaction. Perhaps the eco-plus classes have higher expectations.

# Model Development and Results

## Linear modeling

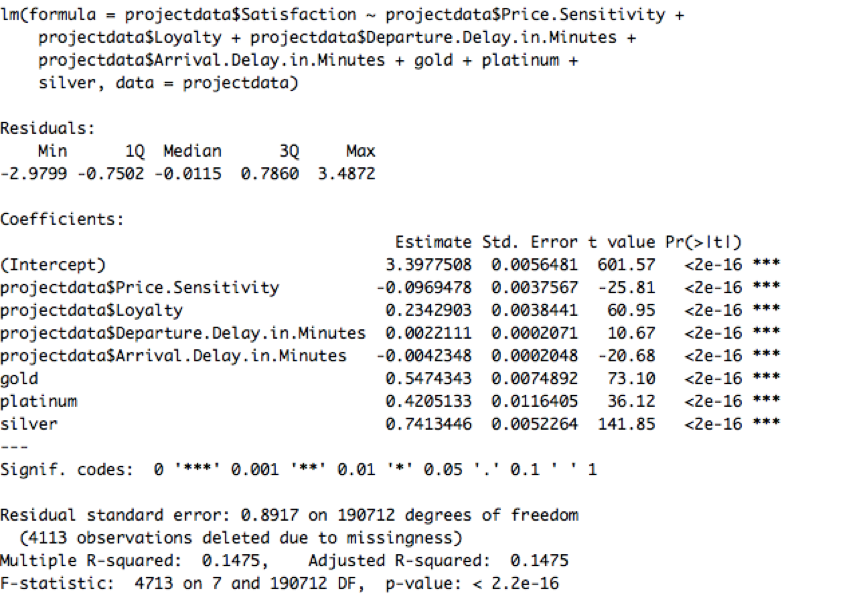
For the qualitative variables, we make them as dummy variables. From the Pearson Correlation Coefficient table, we can see, the correlation coefficient between satisfaction and shopping amount and between satisfaction and eating and drinking amount is pretty low. Thus, we can kick them out from our model.

We use Satisfaction as response variable and loyalty, price sensitivity, departure delay and arrival delay as predictor variables. We also add other qualitative variables into our model such as airline status, type of travel, class and flight cancelled. First, we make these qualitative variables as dummy variables. Next, we built the model. The code is in the appendix as “Linear Model 1 Code” and consequence of the model are shown below.



Summary of a generated linear model for predicting customer satisfaction using loyalty, price sensitivity, departure delay, arrival delay, airline status, type of travel, class and flight cancelled.

From the picture, we can see, the ‘person’ and ‘ecoplus’ and ‘flightcancel’ variables have NA which means multicollinearity between these and other variables and as a result these variables have the same influence on the response variable. Thus, we kick out these variables from our model and built another model called model new. This code is in the appendix, section “Linear Model 2 Code”



Summary of the new linear model for predicting customer satisfaction without the ‘person’ and ‘ecoplus’ and ‘flightcancel’ variables

### Analysis results

From the summary, P-values of all of the variables are all less than 0.05, which means all variables in this model are significant.

When **price sensitivity** increases one unit, the satisfaction will decrease 0.09 unit. Thus, company should attract more customers who have less price sensitivity. Some customers who might have less price sensitivity are those who have higher income or higher social status.

When **loyalty** increases one unit, the satisfaction will increase 0.23 unit. In order to improve satisfaction, company should increase loyalty of customers. This makes sense because loyal customers should be bias towards having higher satisfaction.

When **departure delay** increases one minutes, the satisfaction will increase 0.002 unit. This result is interesting because one would expect the satisfaction to decrease with departure delay. The explanation may be that if departure delay happens, the airline company will give customer some compensation which may increase their satisfaction, or this could be the result of statistical error. (Even we did not include departure delay in all kinds of models, we still would like to analyse it in linear regression model.)

When **arrival delay** increases one minutes, the satisfaction will decrease 0.004 unit. Thus, in order to increase satisfaction, the company should decrease arrival delay.

## Support Vector Machine Modeling

We trained SVM models multiple times using the survey data as training/test data. The exact code is in the appendix as “Support Vector Machine Modelling Code”. It is important to note that we did not generate a SVM model of the whole dataset, as that resulted in the computer freezing up. We attempted to train an SVM model on personal computers and the school computer, but both froze. Instead, the dataset was cut down into a fifteenth of its size and then trained. The six variables with the highest Pearson Correlation Coefficients were used to generate SVM models:

1. Flights Per Year
2. Price Sensitivity
3. Loyalty
4. Arrival Delay in Minutes
5. Departure Delay in Minutes
6. Total Frequent Flyer Accounts

Though Flights Per Year and Price Sensitivity would be hard to affect, it is still important to use them to increase the accuracy of our model. To see the level of effect different variables, have on predicting customer satisfaction, models were trained with two, four, and six variables. The variable of interest (Satisfaction) was split into two categories, happy (where the user selected three or higher in on their survey), and not happy (where the user selected one or two).

### Confusion Matrixes

First a model was trained with just the variables with the top two highest absolute Pearson Correlation Coefficients.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Flights.Per.Year + Loyalty | | |
| Test Data |  | Predicted Not Happy | Predicted Happy |
| Happy | 5 | 5029 |
| Not Happy | 1 | 1323 |

The error rate here was about 21%

Next a model with the variables that had the top four highest absolute Pearson Correlation Coefficients was trained. Below is the confusion matrix for that model:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Flights.Per.Year + Loyalty + Price.Sensitivity + Total.Freq.Flyer.Accts Model | | |
| Test Data |  | Predicted Not Happy | Predicted Happy |
| Happy | 9 | 5025 |
| Not Happy | 7 | 1317 |

Here we see that the SVM model is much better at predicting not happiness than happiness. The error rate here is roughly 21%

Next, a model was trained with the variables that had the top six highest absolute Pearson Correlation Coefficients.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Flights.Per.Year + Loyalty + Price.Sensitivity + Total.Freq.Flyer.Accts + Arrival.Delay.in.Minutes + Departure.Delay.in.Minutes | | |
| Test Data |  | Predicted Not Happy | Predicted Happy |
| Happy | 141 | 4893 |
| Not Happy | 137 | 1187 |

Despite having more variables, the error rate remained steady at 21%

The error rates were relatively the same:

|  |  |
| --- | --- |
| **Predictor Variables** | **Model Error Rate** |
| Two | 21% |
| Four | 21% |
| Six | 21% |

While we could gain some insights through the SVM modelling of bucketed satisfaction variables, we ran into some problems, namely that the SVM model trained solely with the bucketed satisfaction variable and loyalty resulted in a model with just one predicted outcome. This may have been a result of the cut down data set, or a quirk of the data. To verify these rules, it was decided to train a SVM model without bucketing Satisfaction, and see if we get similar results.

### SVM Models without bucketing

Three SVM models were trained regressively without satisfaction bucketing. The first was trained with Flights.Per.Year, Loyalty, Price.Sensitivity, and Total.Freq.Flyer.Accts. The second was trained with all of the first, and Arrival.Delay.in.Minutes added. Finally, the last was trained with Arrival.Delay.in.Minutes and Departure.Delay.in.Minutes added. After each model was trained, the Root Means Squared Error was calculated.

|  |  |
| --- | --- |
|  | Root Means Squared Error |
| Four Predictors | 0.9703008 |
| Five Predictors | 0.9630733 |
| Six Predictors | 0.962188 |

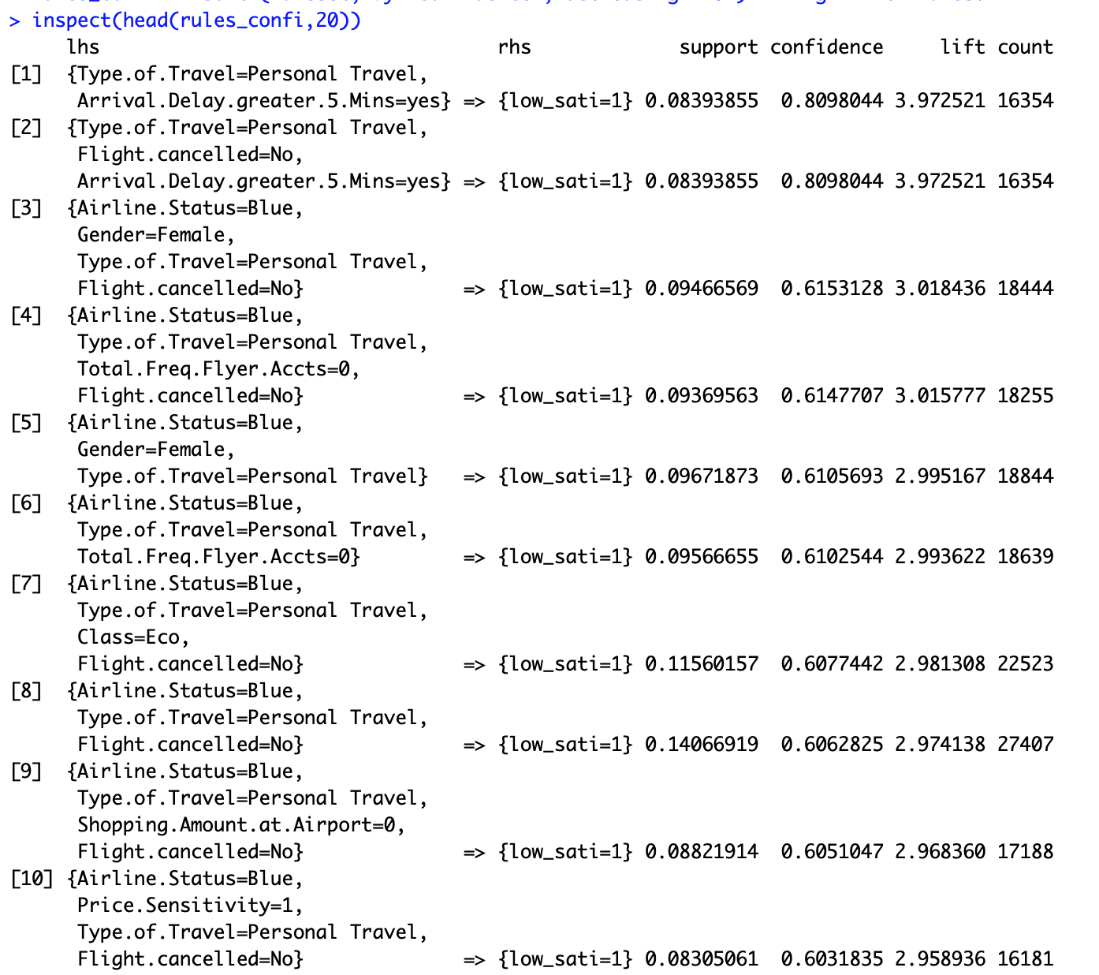
The RMSE confirmed the results shown by the SVM models with bucketing.

### Analysis results

Each of these variables were equally impactful on the customer satisfaction, but addressing every variable produced significant diminishing returns. Based on this trend, the best way to reduce low customer satisfaction is by focusing on one variable rather than all at once. Although Flights Per Year had the highest Pearson Correlation Coefficient, it is relatively unaffectable, so efforts should be focused on the next variable with the highest Pearson Correlation Coefficient, Loyalty.

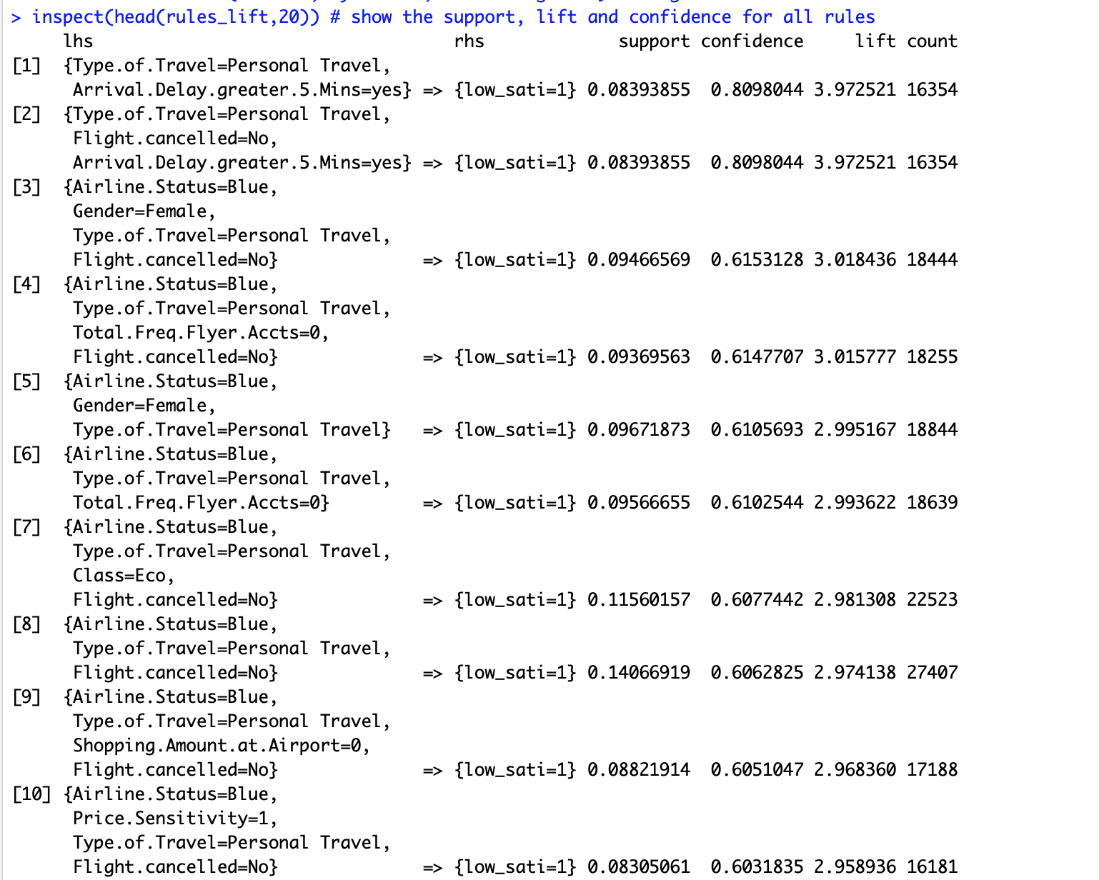
## Association Rule Mining

Due to the linear modeling and SVM modelling we know what some important variables are. We wanted to discover thing outside of our current understandings, so we decided to do association mining. Our hope is that association mining will generate some insights that we do not already know.



The top 15 association rules with the right-hand side equals to low satisfaction, ranked by lift.

There is a great number of “Type of Travel” association rules, and it seems like they are highly correlated “Arrival Delay Greater than 5 mins”. Of particular note is the fact that loyalty did not show up. While loyalty seems associated with high satisfaction it does not seem associated with low satisfaction.



The top 15 association rules with the right-hand side equals to low satisfaction, ranked by lift.

Again, we see the importance of Type of Travel and Arrival Delay. We can see that personal travel, long duration, and arrival delay is highly associated with low customer satisfaction. These effects are especially distinct on female and blue membership customers.

# Actionable Insights and Conclusion

Based on the Pearson Correlation results, we focused on these variables: Flights Per Year, Loyalty, Price Sensitivity, Total Frequent Flyer Miles, and Arrival Delay. We then used association rules, linear regression, and support vector machine modelling to discover which among these variables were has the highest impact on customer satisfaction. The results are below:

1. Association Rules: Personal travel, long duration, and arrival delay
2. Regression: Loyalty, arrival delay and price sensitivity
3. SVM: Flights per Year, loyalty, price sensitivity and arrival delay

We believe variables worth investigated are arrival delay, loyalty, and price sensitivity, due to affectability.

## Reducing Arrival Delay

While arrival delay is not usually directly affected by airline staff, there are still some ways to reduce arrival delay:

* Optimize regular check and maintenance process
* Training of staff
* Using algorithms and historical data to predict flight delays, etc.

## Improving Customer Loyalty

A few ways to improve customer loyalty could be to:

* Make customer service a priority
* Reward your customers
* Ask customers for advice.

For example, making sure to always promptly answer questions, even online (Baird, 2011) (Furgison, 2017). Spending resources to make sure social media accounts are friendly and engaging can yield results, but companies must be careful to not confuse their own desire for customer intimacy and what the consumers really want.

## Reducing the effects of Price Sensitivity

Research has shown better brand credibility can reduce the negative effects of price sensitivity (Tülin Erdem, 2002).

* Airlines could start a marketing campaigned aimed towards increasing brand credibility to indirectly alleviate price sensitivity issues.
* Another way to increase brand credibility is to gather and publicize customer testimonials (Drucker, 2018).

In conclusion, the customer dissatisfaction is caused by a composite of multiple factors. Important factors that can be affected are customer loyalty, arrival delay, and price sensitivity. Customer loyalty has a positive effect on customer satisfaction, while price sensitivity and arrival delay lower customer satisfaction. Airlines should increase responsiveness on social media accounts to improve customer loyalty. Airlines should also improve brand credibility to suppress the negative effects of price sensitivity. And airlines should also decline the possibility of arrival delay by optimizing regular check and maintenance process, training of staff and using algorithms and historical data to predict flight delays.

# Appendix

## Airline Status, Types of Travel, Class of Flight Code

> tapply(fdf$Satisfaction\_Num, fdf$Airline.Status, mean, na.rm=TRUE)

Blue Gold Platinum Silver

3.159800 3.750841 3.621126 3.945951

> tapply(fdf$Satisfaction\_Num, fdf$Type.of.Travel, mean, na.rm=TRUE)

Business travel Mileage tickets Personal Travel

3.775223 3.555723 2.554625

> tapply(fdf$Satisfaction\_Num, fdf$Class, mean, na.rm=TRUE)

Business Eco Eco Plus

3.543016 3.373626 3.314364

## Pearson Correlation Coefficient Code

# Run these three functions to get a clean test of code

dev.off() # Clear the graph window

cat('\014') # Clear the console

rm(list=ls()) # Clear all user objects from the environment!!!

# Set working directory

# Change to the folder containing your data files

setwd("/Users/liije/Documents/Spring 2019/IST 687 Data Science")

library(kernlab)

csvToRead <- "spring19survey.csv"

surveyData <- read.csv(csvToRead, stringsAsFactors = FALSE)

set.seed(8081)

str(surveyData)

Candidate.Variable <- c("Flights Per Year", "Loyalty", "Total Freq Flyer Accts", "Scheduled Dep Hour", "Departure Delay", "Arrival Delay", "Flight Time", "Price Sensitivity", "Shopping Amount", "Eating Amount")

Pearson.Correlation.Coefficient <- c(

cor.test(surveyData$Flights.Per.Year, surveyData$Satisfaction)$estimate,

cor.test(surveyData$Loyalty, surveyData$Satisfaction)$estimate,

cor.test(surveyData$Total.Freq.Flyer.Accts, surveyData$Satisfaction)$estimate,

cor.test(surveyData$Scheduled.Departure.Hour, surveyData$Satisfaction)$estimate,

cor.test(surveyData$Departure.Delay.in.Minutes, surveyData$Satisfaction)$estimate,

cor.test(surveyData$Arrival.Delay.in.Minutes, surveyData$Satisfaction)$estimate,

cor.test(surveyData$Flight.time.in.minutes, surveyData$Satisfaction)$estimate,

cor.test(surveyData$Price.Sensitivity, surveyData$Satisfaction)$estimate,

cor.test(surveyData$Shopping.Amount.at.Airport, surveyData$Satisfaction)$estimate,

cor.test(surveyData$Eating.and.Drinking.at.Airport, surveyData$Satisfaction)$estimate)

pearsonCorrelation <- data.frame(Candidate.Variable, Pearson.Correlation.Coefficient)

View(pearsonCorrelation)

## SVM Modeling Code (In R)

# Run these three functions to get a clean test of code

dev.off() # Clear the graph window

cat('\014') # Clear the console

rm(list=ls()) # Clear all user objects from the environment!!!

# Set working directory

# Change to the folder containing your data files

setwd("/Users/liije/Documents/Spring 2019/IST 687 Data Science")

library(kernlab)

csvToRead <- "spring19survey.csv"

surveyData <- read.csv(csvToRead, stringsAsFactors = FALSE)

surveyData <- surveyData[!is.na(surveyData$Satisfaction), ]

surveyData <- surveyData[!is.na(surveyData$Departure.Delay.in.Minutes), ]

surveyData <- surveyData[!is.na(surveyData$Arrival.Delay.in.Minutes), ]

surveyData <- surveyData[!is.na(surveyData$Loyalty), ]

surveyData <- surveyData[!is.na(surveyData$Total.Freq.Flyer.Accts), ]

# Cut the survey Data into a twentieth of it's size.

cutPoint1\_10 <- floor(dim(surveyData)[1]/10)

surveyData <- surveyData[1:cutPoint1\_10,]

#Put customer satisfaction into buckets

vBuckets <- replicate(length(surveyData$Satisfaction), "happy")

vBuckets[surveyData$Satisfaction < 3] <- "notHappy"

surveyData$Satisfaction.Bucket <- as.factor(vBuckets)

# Split survey data into test and training data (two thirds is training, one third is test)

randIndex <- sample(1:dim(surveyData)[1])

cutPoint2\_3 <- floor(2 \* dim(surveyData)[1]/3)

trainData <- surveyData[randIndex[1:cutPoint2\_3],]

testData <- surveyData[randIndex[(cutPoint2\_3+1):dim(surveyData)[1]],]

# Arrival SVM Model

Arrival.Delay.in.Minutes.SvmOutput <- ksvm(Satisfaction.Bucket ~ Arrival.Delay.in.Minutes, data=trainData, kernel="rbfdot", kpar="automatic",C=5,cross=3, prob.model=TRUE)

# Look at model

Arrival.Delay.in.Minutes.SvmPred <- predict(Arrival.Delay.in.Minutes.SvmOutput, testData, type = "votes")

Arrival.Delay.in.Minutes.CompTable <- data.frame(testData[,30],Arrival.Delay.in.Minutes.SvmPred[1,])

table(Arrival.Delay.in.Minutes.CompTable)

(17+650)/(17+2506+6+650)

(15+832)/(3379+15+832+12)

(4+1340)/(4+5001+3+1350)

# Train Loyalty SVM Model

Loyalty.SvmOutput <- ksvm(Satisfaction.Bucket ~ Loyalty, data=trainData, kernel="rbfdot", kpar="automatic",C=5,cross=3, prob.model=TRUE)

# Look at model

Loyalty.SvmPred <- predict(Loyalty.SvmOutput, testData, type = "votes")

Loyalty.CompTable <- data.frame(testData[,30],Loyalty.SvmPred[1,])

table(Loyalty.CompTable)

(5005+1353)

# Train Loyalty & Arrival SVM Model

Loyalty.Arrival.SvmOutput <- ksvm(Satisfaction.Bucket ~ Loyalty + Arrival.Delay.in.Minutes, data=trainData, kernel="rbfdot", kpar="automatic",C=5,cross=3, prob.model=TRUE)

# Look at model

Loyalty.Arrival.SvmPred <- predict(Loyalty.Arrival.SvmOutput, testData, type = "votes")

Loyalty.Arrival.CompTable <- data.frame(testData[,30],Loyalty.Arrival.SvmPred[1,])

table(Loyalty.Arrival.CompTable)

(55+1308)/(55+1308+45+4950)

# Train Departure.Delay & Total.Freq. SVM Model

Dept.Freq.SvmOutput <- ksvm(Satisfaction.Bucket ~ Departure.Delay.in.Minutes + Total.Freq.Flyer.Accts, data=trainData, kernel="rbfdot", kpar="automatic",C=5,cross=3, prob.model=TRUE)

# Look at model

Dept.Freq.SvmPred <- predict(Dept.Freq.SvmOutput, testData, type = "votes")

Dept.Freq.CompTable <- data.frame(testData[,30], Dept.Freq.SvmPred[1,])

table(Dept.Freq.CompTable)

(30+629)/(36+2488+26+629)

(27+672)/(27+672+2462+18)

( 6 +1347)/(6+4999+6+1347)

# Train 4 variable SVM Model

fourVarSvmOutput <- ksvm(Satisfaction.Bucket ~ Loyalty + Arrival.Delay.in.Minutes + Departure.Delay.in.Minutes + Total.Freq.Flyer.Accts, data=trainData, kernel="rbfdot", kpar="automatic",C=5,cross=3, prob.model=TRUE)

# Look at model

fourVarSvmPred <- predict(fourVarSvmOutput, testData, type = "votes")

fourVarCompTable <- data.frame(testData[,30],fourVarSvmPred[1,])

table(fourVarCompTable)

# Calculate error

(60+622)/(60+33+2464+622)

(49+654)/(49+2440+652+36)

(103+760)/(84+760+3291+103)

# Train 2 pred 2 UnPredict variable SVM Model

twoPredTwoUnpredictSvmOutput <- ksvm(Satisfaction.Bucket ~ Loyalty + Flights.Per.Year + Price.Sensitivity + Total.Freq.Flyer.Accts, data=trainData, kernel="rbfdot", kpar="automatic",C=5,cross=3, prob.model=TRUE)

# Look at model

twoPredTwoUnpredictSvmPred <- predict(twoPredTwoUnpredictSvmOutput, testData, type = "votes")

twoPredTwoUnpredictCompTable <- data.frame(testData[,30],twoPredTwoUnpredictSvmPred[1,])

table(twoPredTwoUnpredictCompTable)

# Calculate error

(9+1317)/(9+1317+5025+7)

# Train 4 pred 2 UnPredict SVM Model

fourPredTwoUnpredictSvmOutput <- ksvm(Satisfaction.Bucket ~ Loyalty + Flights.Per.Year + Price.Sensitivity + Total.Freq.Flyer.Accts + Arrival.Delay.in.Minutes + Departure.Delay.in.Minutes, data=trainData, kernel="rbfdot", kpar="automatic",C=5,cross=3, prob.model=TRUE)

# Look at model

fourPredTwoUnpredictSvmPred <- predict(fourPredTwoUnpredictSvmOutput, testData, type = "votes")

fourPredTwoUnpredictCompTable <- data.frame(testData[,30],fourPredTwoUnpredictSvmPred[1,])

table(fourPredTwoUnpredictCompTable)

# Calculate error

(141+1187)/(141+4893+137+1187)

# Train 1 pred 1 UnPredict SVM Model

onePredOneUnPredictSvmOutput <- ksvm(Satisfaction.Bucket ~ Flights.Per.Year + Loyalty, data=trainData, kernel="rbfdot", kpar="automatic",C=5,cross=3, prob.model=TRUE)

# Look at model

onePredOneUnPredictSvmPred <- predict(onePredOneUnPredictSvmOutput, testData, type = "votes")

onePredOneUnPredictCompTable <- data.frame(testData[,30],onePredOneUnPredictSvmPred[1,])

table(onePredOneUnPredictCompTable)

# Calculate error

(5+1323)/(5+1323+1+5029)

## Linear Model 2 Code

modelnew<-lm(projectdata$Satisfaction~projectdata$Price.Sensitivity+projectdata$Loyalty+projectdata$Departure.Delay.in.Minutes+projectdata$Arrival.Delay.in.Minutes+gold+platinum+silver,data = projectdata)

summary(modelnew)

## Satisfaction Data Discovery Code

# Run these three functions to get a clean test of code

dev.off() # Clear the graph window

cat('\014') # Clear the console

rm(list=ls()) # Clear all user objects from the environment!!!

# Set working directory

# Change to the folder containing your data files

setwd("/Users/liije/Documents/Spring 2019/IST 687 Data Science")

library(ggplot2)

csvToRead <- "spring19survey.csv"

surveyData <- read.csv(csvToRead, stringsAsFactors = FALSE)

satisfaction.vec <- surveyData[!is.na(surveyData$Satisfaction), ]

satisfaction.vec <- satisfaction.vec$Satisfaction

summary(satisfaction.vec)

quantile(satisfaction.vec, c(0.05, 0.95))

## Association Rule Code

dev.off() # Clear the graph window

cat('\014') # Clear the console

rm(list=ls()) # Clear all user objects from the environment!!!

fdf <- read.csv('/Users/dahailiu/Desktop/spring19survey.csv')

### Define response variable

#### Association Rule Starts Here

## Prepare dataset

library(arules)

library(arulesViz)

fdf$low\_sati <- ifelse(fdf$Satisfaction >2, '0' , '1')

fdf\_clean\_df <- subset(fdf, select = -c(Flight.date, Partner.Name,Orgin.City,Destination.City,Scheduled.Departure.Hour,Departure.Delay.in.Minutes,

Arrival.Delay.in.Minutes,Flight.time.in.minutes, Satisfaction) )

fdft\_trans <- as(data.frame(lapply(fdf\_clean\_df, as.factor), stringsAsFactors=T), "transactions")

size(head(fdft\_trans, 3))

inspect(head(fdft\_trans, 3))

## Explore data in Rules:

itemFrequencyPlot(fdft\_trans, support = 0.2, topN = 20, type="absolute")

ruleset <- apriori(fdft\_trans, parameter = list(support = 0.08, confidence = 0.3), appearance = list (default="lhs",rhs="low\_sati=1"))

summary(ruleset)

inspect(head(ruleset,15))

rules\_lift <- sort (ruleset, by="lift", decreasing=TRUE) # 'high-lift' rules.

inspect(head(rules\_lift,20)) # show the support, lift and confidence for all rules

rules\_confi <- sort (ruleset, by="confidence", decreasing=TRUE) # 'high-lift' rules.

inspect(head(rules\_confi,20))

prop.table(table(fdf\_clean\_df$Type.of.Travel, fdf\_clean\_df$low\_sati))

library(arulesViz)

plot(ruleset)

goodrules <- ruleset[quality(ruleset)$lif>3]

inspect(goodrules)

## SVM Code without Bucketing

# Run these three functions to get a clean test of code

dev.off() # Clear the graph window

cat('\014') # Clear the console

rm(list=ls()) # Clear all user objects from the environment!!!

# Set working directory

# Change to the folder containing your data files

setwd("/Users/liije/Documents/Spring 2019/IST 687 Data Science")

library(kernlab)

csvToRead <- "spring19survey.csv"

surveyData <- read.csv(csvToRead, stringsAsFactors = FALSE)

surveyData <- surveyData[!is.na(surveyData$Satisfaction), ]

surveyData <- surveyData[!is.na(surveyData$Flights.Per.Year), ]

surveyData <- surveyData[!is.na(surveyData$Loyalty), ]

surveyData <- surveyData[!is.na(surveyData$Price.Sensitivity), ]

surveyData <- surveyData[!is.na(surveyData$Arrival.Delay.in.Minutes), ]

# Cut the survey Data down

cutPoint1\_10 <- floor(dim(surveyData)[1]/15)

surveyData <- surveyData[1:cutPoint1\_10,]

# Split survey data into test and training data (two thirds is training, one third is test)

randIndex <- sample(1:dim(surveyData)[1])

cutPoint2\_3 <- floor(2 \* dim(surveyData)[1]/3)

trainData <- surveyData[randIndex[1:cutPoint2\_3],]

testData <- surveyData[randIndex[(cutPoint2\_3+1):dim(surveyData)[1]],]

# Train four predictor model

FourModelSvmOutput <- ksvm(Satisfaction ~ Flights.Per.Year + Loyalty + Price.Sensitivity + Total.Freq.Flyer.Accts, data=trainData, kernel="rbfdot", kpar="automatic",C=1,cross=3, prob.model=TRUE)

FourModelSvmPred <- predict(FourModelSvmOutput, testData, type = "response")

sqrt(mean((testData$Satisfaction-FourModelSvmPred)^2))

#Satisfaction minus testData

# Train five predictor model

FiveModelSvmOutput <- ksvm(Satisfaction ~ Flights.Per.Year + Loyalty + Price.Sensitivity + Total.Freq.Flyer.Accts + Arrival.Delay.in.Minutes, data=trainData, kernel="rbfdot", kpar="automatic",C=1,cross=3, prob.model=TRUE)

FiveModelSvmPred <- predict(FiveModelSvmOutput, testData, type = "response")

sqrt(mean((testData$Satisfaction-FiveModelSvmPred)^2))

# Train six predictor model

SixModelSvmOutput <- ksvm(Satisfaction ~ Flights.Per.Year + Loyalty + Price.Sensitivity + Total.Freq.Flyer.Accts + Arrival.Delay.in.Minutes + Departure.Delay.in.Minutes, data=trainData, kernel="rbfdot", kpar="automatic",C=1,cross=3, prob.model=TRUE)

SixModelSvmPred <- predict(SixModelSvmOutput, testData, type = "response")

sqrt(mean((testData$Satisfaction-FiveModelSvmPred)^2))

# Bibliography

Baird, C. H. (2011). From social media to social customer relationship management. *Strategy & Leadership*, 30-37.

Drucker, J. (2018, May 15). *The Importance Of Brand Credibility And How To Build It*. Retrieved from Forbes: https://www.forbes.com/sites/theyec/2018/05/15/the-importance-of-brand-credibility-and-how-to-build-it/#7845eefa26fd

Furgison, L. (2017). *5 Ways to Increase Customer Loyalty*. Retrieved from Five Stars: https://blog.fivestars.com/5-ways-to-increase-customer-loyalty/

Tülin Erdem, J. S. (2002). The impact of brand credibility on consumer price sensitivity. *International Journal of Research in Marketing Volume 19, Issue 1*, 1-19.