

Airline Customer Satisfaction

ISt 687 – Applied Data science Prediction

Shloak Gupta || Zhenyu Shao || Wenjin Yao || Yu Yu Ko || Satyen Amonkar

# Appendix

|  |  |
| --- | --- |
| Introduction | 2 |
| Business Questions | 2 |
| Data Wrangling & Munging | 3 |
| Visualizations | 7 |
| Stepwise Regression | 11 |
| Linear Regression | 19 |
| Association Rules | 22 |
| Decision tree | 29 |
| Support Vector Machine | 34 |
| Insights and Conclusions | 36 |

# Introduction

This project aims to analyze airline customer satisfaction data of the flights in the United States of America and provide actionable insights that will enable them to improve satisfaction. Since the economics of the industry is majorly dependent on customer satisfaction, it is essential to get insights into factors affecting it.

The dataset provided is huge around 130,000 rows and 28 variables. The Customer Ratings are on the scale 0-5. The Data is from 14 different airlines flying in the USA. Finding the significant factors that affect the rating and how do they do it is the primary question which leads to other subqueries.

The general approach followed through this Project is:

* Brainstorming for the pertinent data questions.
* Cleaning the data for irregularities and replacing missing information
* Identifying a set of relevant variables
* Finding relation between variables and to the satisfaction rating
* Creating prediction models to show performance of these variables
* Validating performance among different models
* Providing actionable insights after the analysis

# Business Questions

* What are the major attributes affecting the ratings?
* What are the reasons for high ratings?
* What are the reasons for low ratings?
* What can be done to avoid low ratings?
* Is gender, type of travel, Airline Status, Flight Distance are affecting Ratings?
* Arrival delay, Flight Cancellation, Price Sensitivity are conceived to be important factors. Are they that important?
* Do factors like Eating and Shopping which are not directly associated with the flight affect rating?
* What are the best models and variables to predict rating of customer?

# Data Wrangling and Munging

As with every realistic data this data also had problem there is always an option of deleting the misbehaving part of the data. But as it is been observed from past that every data is important, and we should correct theses parts. Major time is this project as with all data science projects a lot of time was spend on framing the data for prediction analysis. Data pre-processing

1. The satisfaction column in the given data set was “character” data type which had to be converted to the numeric data type by correcting typing mistakes in three rows of the data.
2. Day to the week might affect the ratings more than the day or the month of the year. To test this day of the month and the Flight Date columns were used to make a column weekday.

*survey$Month <- paste ("0”, as.character (survey$Month), sep="")*

*survey$date\_of\_flight <- paste (as.character (survey$Year), as.character(survey$Month),*

*as.character (survey$Day.of.Month), sep = "-")*

*#View(survey)*

*survey$weekday <- weekdays (as.Date (survey$date\_of\_flight))*

*survey <- subset (survey, select =-c (Month, Year))*

For Example:

Flight Date: 3/18/14 is converted to Weekday : Tuesday

1. Missing values in Arrival Delay and Departure Delay Columns

* It was observed that the major reason for the missing values is when the flight is cancelled. Removing all the rows with missing value is not a proper alternative in this case.
* Along with this three columns Arrival Delay, Flight Cancelled and arrival delay Greater than 5 mins can be converted to a single column with no missing values.
* Same is with departure delay and Departure delay Greater than 5 mins.

*arrivaldel <- function(vec1, vec){*

*vBuckets <- replicate(length(vec),"No Delay")*

***vBuckets[is.na(vec) & vec1 == 'Yes'] <- "Cancelled"***

*vBuckets[vec > 0 & vec < 5] <- "5 minutes"*

*vBuckets[vec >= 5 & vec < 30] <- "half hours"*

*vBuckets[vec >= 30 & vec < 60] <- "1 hours"*

*vBuckets[vec >= 60 & vec < 120] <- "2 hours"*

*vBuckets[vec >= 120] <- "more than 2 hours"*

*return (vBuckets)*

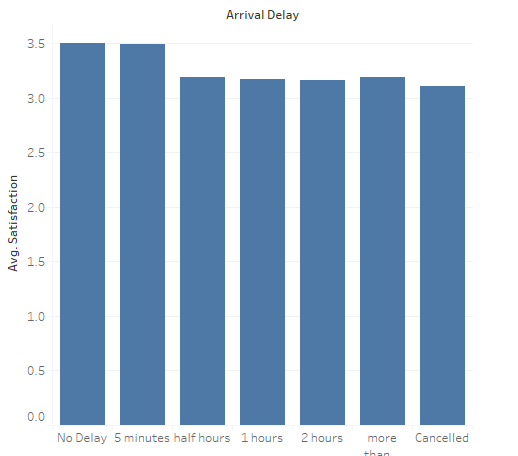
*}*

*survey$arrivalDelay <-arrivaldel(survey$Flight.cancelled, survey$Arrival.Delay.in.Minutes)*

*survey$DepartureDelay <- arrivaldel(survey$Flight.cancelled,survey$Departure.Delay.in.Minutes)*

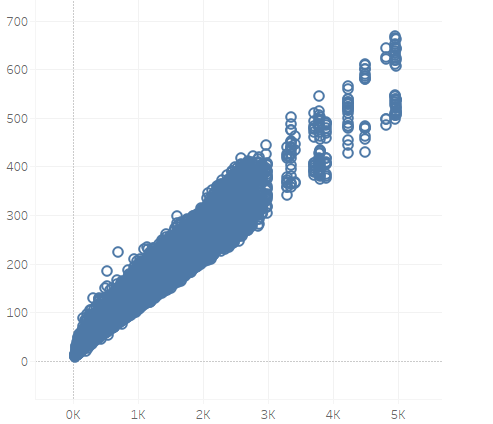
*survey <- subset(survey, select =-c(Arrival.Delay.greater.5.Mins))*

*survey <- subset(survey, select =-c(Arrival.Delay.in.Minutes,Departure.Delay.in.Minutes))*

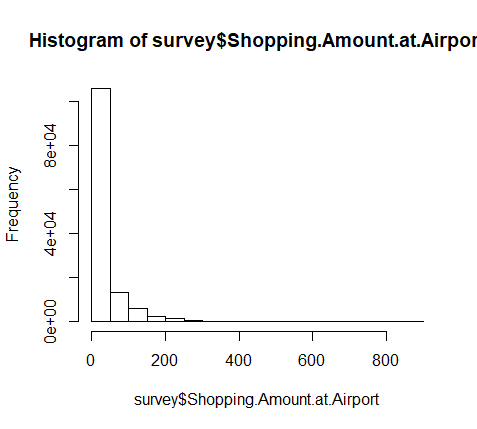
**

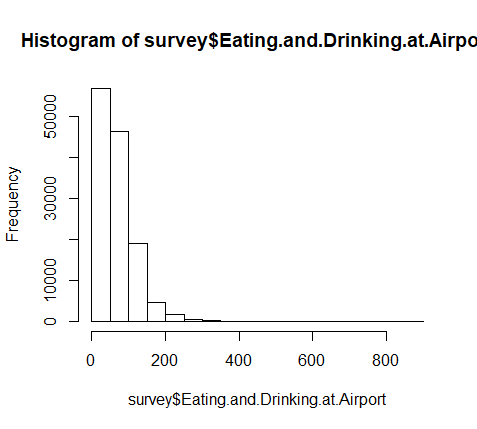
1. The missing values were also present in the Flight Time in minutes column of the dataset. It was found out that Flight Distance is highly correlated to Flight Time so we dropped Flight Time.

*cor(survey$Flight.time.in.minutes, survey$Flight.Distance).*



1. Two columns in Shopping amount at airport and Eating and Drinking at airport are very right skewed data so they were changed to categorical variables.

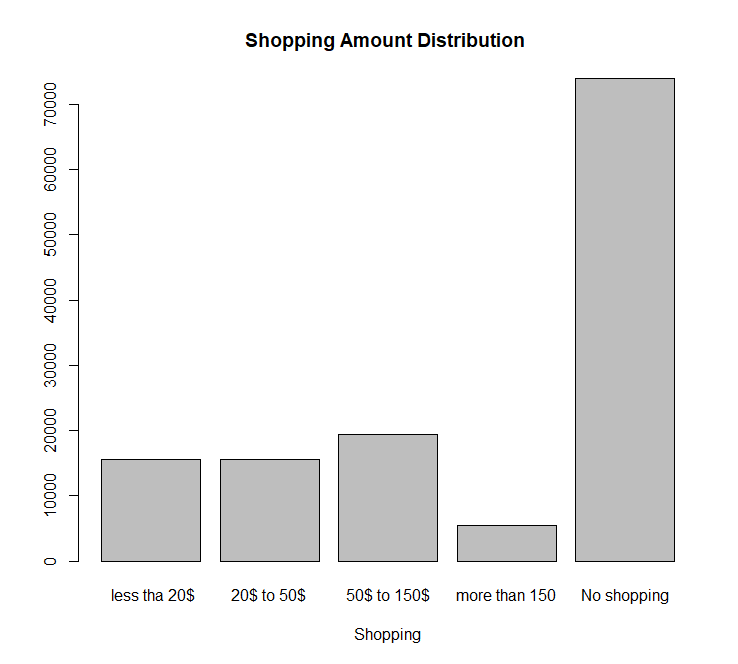




After changing them to the categorical variables just as we did for delay column.

*counts <- table(survey$shop\_at\_airport)*

*barplot(counts, xlab = 'Shopping', main = 'Shopping Amount Distribution ')*



1. Airline Name and Airline Code represent the same thing in the dataset that is a unique company, so we removed the Airline Code column from the Dataset.
2. How old an Airplane is might affect the rating of the customer, so we used the Year of First Flight column and subtracted each row from 2012 as it was the latest year and we got the column Years Old.

# Visualizations

*surveyVisualizationData<-read.csv("surveyDataV7.csv")*

*surveyVisualizationData<- data.frame(surveyVisualizationData)*

*library(ggplot2)*

*library(sqldf)*

*surveyVisualizationData$TOT<-surveyVisualizationData$Type.of.Travel*

*surveyVisualizationData$Type.of.Travel*

*y1<-sqldf('select AVG(Satisfaction) as AVG\_Satisfaction,AirlineStatus,TOT from surveyVisualizationData Group by AirlineStatus, TOT')*

*x<-ggplot(y1,aes(x=TOT, y=AVG\_Satisfaction, fill = factor(AirlineStatus), bgcolor="Black"))*

*x<-x+ geom\_bar(stat="identity", position="dodge")*

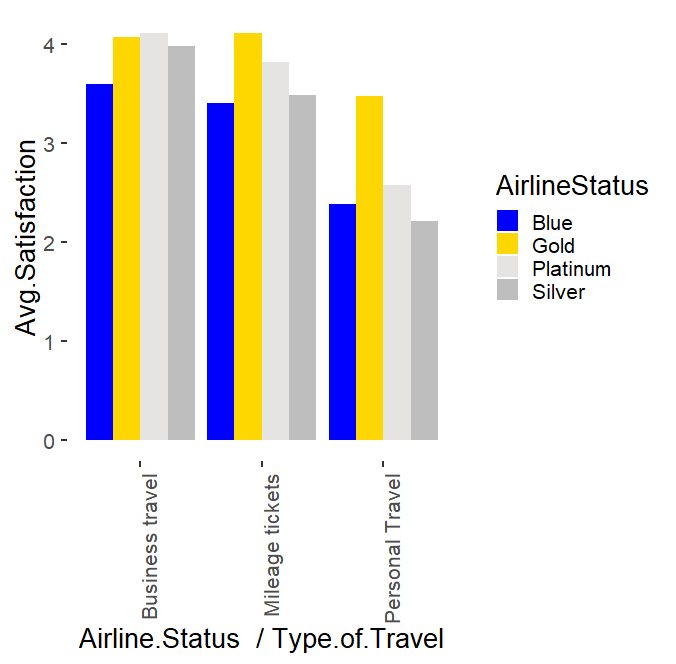
*x<-x+ scale\_fill\_manual(name="AirlineStatus",breaks=c(1,2,3,4), labels=c("Blue","Gold","Platinum","Silver"), values = c("Blue","Gold", '#e5e4e2',"Grey"))*

*x<-x+theme(panel.background = element\_rect(fill = "white",colour = "white",size = 0.5, linetype = "solid"),panel.grid.major = element\_line(size = 0.5, linetype = 'solid',colour = "white"),panel.grid.minor = element\_line(size = 0.25, linetype = 'solid',colour = "white"))*

*x<- x+xlab("Airline.Status / Type.of.Travel")+ylab("Avg.Satisfaction")*

*x<- x+ theme(axis.text.x = element\_text(angle = 90, hjust = 1))*

*x*

**

*surveyVisualizationData<-read.csv("surveyDataV7.csv")*

*surveyVisualizationData<- data.frame(surveyVisualizationData)*

*surveyVisualizationData$Flight\_cancelled<-surveyVisualizationData$Flight.cancelled*

*surveyVisualizationData$Price\_Sensitivity<-surveyVisualizationData$Price.Sensitivity*

*library(sqldf)*

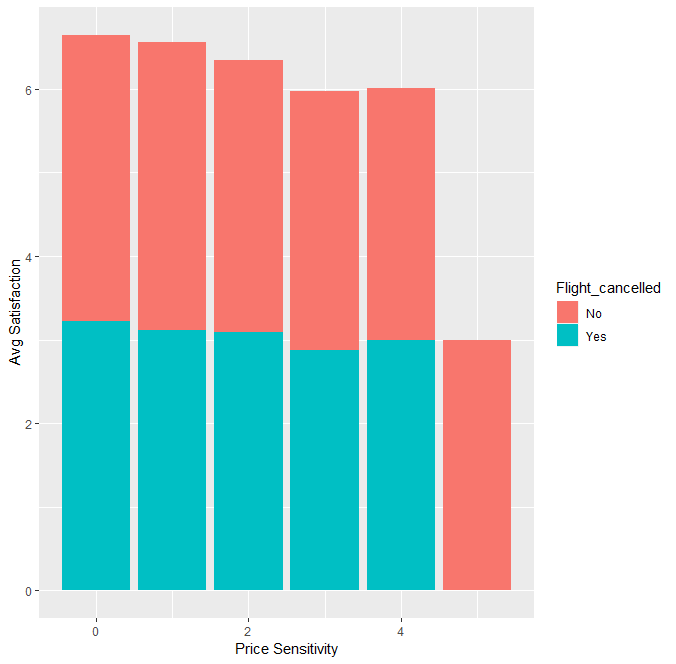
*y<-sqldf('select AVG(Satisfaction) as AVG\_Satisfaction,Price\_Sensitivity,Flight\_cancelled from surveyVisualizationData Group by Price\_Sensitivity, Flight\_cancelled')*

*library(ggplot2)*

*x<-ggplot(y,aes(x=Price\_Sensitivity, y=AVG\_Satisfaction, fill=Flight\_cancelled))*

*x<- x+geom\_col() +xlab("Price Sensitivity")+ylab("Avg Satisfaction")*

*x*

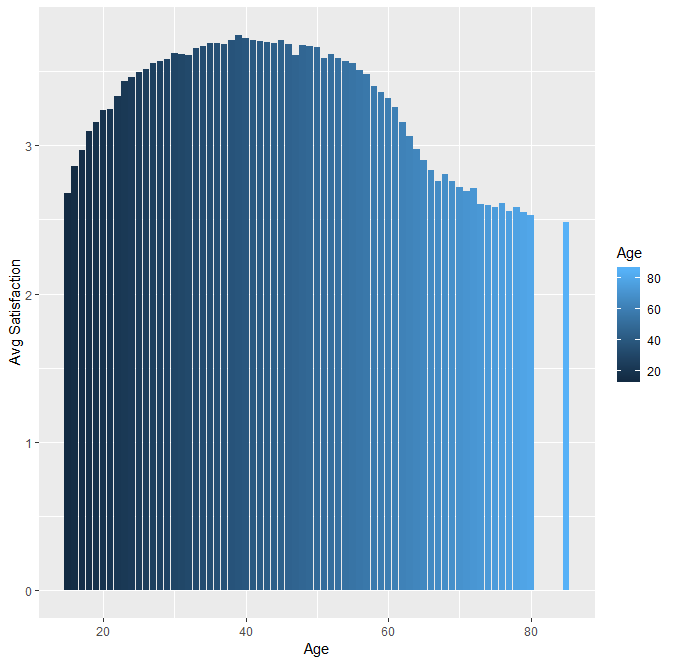
**

*y2<-sqldf('select AVG(Satisfaction) as AVG\_Satisfaction,Age from surveyVisualizationData Group by Age')*

*x2<-ggplot(y2,aes(x=Age, y=AVG\_Satisfaction, fill=Age))*

*x2<- x2+geom\_col() +xlab("Age")+ylab("Avg Satisfaction")*

*x2*



# Stepwise Regression

Stepwise Regression is to add and remove potential influence factors to the regression model. This model is meaningful especially the dataset contains a lot of dependent factors. By the Adjusted R-Squared, we could evaluate a factor is important or not.

*# Importing the dataset*

*setwd("C:/Users/jason/Desktop/jason/syracuse/IST687/project")*

*dataset = read.csv('surveyDataV7\_yuyu.csv')*

*#dataset = dataset[, 2:15]*

*#Age Adjusted R-squared: 0.04921*

*regressor = lm(formula = Satisfaction ~ Age,*

*data = dataset)*

*summary(regressor)*

A screenshot of a computer

Description generated with very high confidence

1. Age+Gender Adjusted R-squared: 0.06447

*regressor = lm(formula = Satisfaction ~ Age + Gender\_cat,*

*data = dataset)*

*summary(regressor)*

A screenshot of a cell phone

Description generated with very high confidence

1. Age+ Gender + Price.Sensitivity Adjusted R-squared: 0.07449

*regressor = lm(formula = Satisfaction ~ Age + Gender\_cat +Price.Sensitivity,*

*data = dataset)*

*summary(regressor)*

A screenshot of a computer

Description generated with very high confidence

1. Age+Gender + Price.Sensitivity + Years.old Adjusted R-squared: 0.07482

*regressor = lm(formula = Satisfaction ~ Age + Gender\_cat + Price.Sensitivity+ Years.Old,*

*data = dataset)*

*summary(regressor)*

A screenshot of a computer

Description generated with very high confidence

1. Age+Gender + Price.Sensitivity + Years.old + No.of.Flights.p.a. Adjusted R-squared: 0.1139

*regressor = lm(formula = Satisfaction ~ Age + Gender\_cat +Price.Sensitivity+Years.Old+No.of.Flights.p.a.,*

*data = dataset)*

*summary(regressor)*

A screenshot of a computer

Description generated with very high confidence

1. Age+Gender + Price.Sensitivity + Years.old + No.of.Flights.p.a.+Type\_travel\_cat Adjusted R-squared: 0.337

*regressor = lm(formula = Satisfaction ~ Age + Gender\_cat +Price.Sensitivity+Years.Old+No.of.Flights.p.a.*

*+ Type\_travel\_cat,*

*data = dataset)*

*summary(regressor)*

A screen shot of a computer

Description generated with very high confidence

1. Age+Gender + Price.Sensitivity + Years.old + No.of.Flights.p.a.+Type\_travel\_cat+Class Adjusted R-squared: 0.3373

*regressor = lm(formula = Satisfaction ~ Age + Gender\_cat + Price.Sensitivity+ Years.Old+No.of.Flights.p.a.*

*+ Type\_travel\_cat + Class,*

*data = dataset)*

*summary(regressor)*

A screen shot of a computer

Description generated with very high confidence

1. Age+Gender + Price.Sensitivity + Years.old + No.of.Flights.p.a.+Type\_travel\_cat+Class + Scheduled.Departure.Hour Adjusted R-squared: 0.3373
2. Scheduled.Departure.Hour not significant.

*regressor = lm(formula = Satisfaction ~ Age + Gender\_cat +Price.Sensitivity+Years.Old+No.of.Flights.p.a.*

*+ Type\_travel\_cat + Class + Scheduled.Departure.Hour,*

*data = dataset)*

*summary(regressor)*

A screen shot of a computer

Description generated with very high confidence

1. Age+Gender + Price.Sensitivity + Years.old + No.of.Flights.p.a.+Type\_travel\_cat+Class + Scheduled.Departure.Hour + arrivalDelay Adjusted R-squared: 0.3549

*regressor = lm(formula = Satisfaction ~ Age + Gender\_cat +Price.Sensitivity+Years.Old+No.of.Flights.p.a.*

*+ Type\_travel\_cat + Class + arrivalDelay,*

*data = dataset)*

*summary(regressor)*

A screen shot of a computer

Description generated with very high confidence

1. Age+Gender + Price.Sensitivity + Years.old + No.of.Flights.p.a.+Type\_travel\_cat+Class + arrivalDelay +DepartureDelay Adjusted R-squared: 0.3555

*regressor = lm(formula = Satisfaction ~ Age + Gender\_cat +Price.Sensitivity+Years.Old+No.of.Flights.p.a.*

*+ Type\_travel\_cat + Class + arrivalDelay + DepartureDelay,*

*data = dataset)*

*summary(regressor)*

A screen shot of a computer

Description generated with very high confidence

1. Age+Gender + Price.Sensitivity + Years.old + No.of.Flights.p.a.+Type\_travel\_cat+Class + arrivalDelay + DepartureDelay +shop\_at\_airport Adjusted R-squared: 0.3559

*regressor = lm(formula = Satisfaction ~ Age + Gender\_cat +Price.Sensitivity+Years.Old+No.of.Flights.p.a.*

*+ Type\_travel\_cat + Class + arrivalDelay + DepartureDelay + shop\_at\_airport,*

data = dataset)

summary(regressor)

A screen shot of a computer

Description generated with very high confidence

* For the multiple regression, we tried the machine learning to build a prediction model.
* For the prediction model, we did correlation accuracy and found the accuracy is 0.6223. The head() and tail() function could help us compare the result briefly.
* Moreover, we used min\_max accuracy method and got the result of **0.8383.**

*# Importing the dataset*

*setwd("C:/Users/jason/Desktop/jason/syracuse/IST687/project")*

*dataset = read.csv('surveyDataV7\_yuyu.csv')*

*#dataset = dataset[, 2:15]*

*# Fitting Linear Regression to the dataset*

*regressor = lm(formula = Satisfaction ~.,*

*data = dataset)*

*summary(regressor)*

*#delete not signifi. variables*

*dataset$No..of.other.Loyalty.Cards <- NULL*

*dataset$Flight.Distance <- NULL*

*#Splitting the dataset into the Training set and Test set*

*#install.packages('caTools')*

*library(caTools)*

*set.seed(123)*

*split = sample.split(dataset$Satisfaction, SplitRatio = 0.8)*

*training\_set = subset(dataset, split == TRUE)*

*test\_set = subset(dataset, split == FALSE)*

*regressor = lm(formula = Satisfaction ~.,*

*data = training\_set)*

*summary(regressor)*

*# Feature Scaling*

*# training\_set[, 2:3] = scale(training\_set[, 2:3])*

*# test\_set[, 2:3] = scale(test\_set[, 2:3])*

*y\_pred = predict(regressor, newdata = test\_set)*

*y\_pred*

*actuals\_preds <- data.frame(cbind(actuals=test\_set$Satisfaction, predicteds=y\_pred)) # make actuals\_predicteds dataframe.*

*correlation\_accuracy <- cor(actuals\_preds)*

*correlation\_accuracy*

A close up of a sign

Description generated with high confidence

head(actuals\_preds)

A close up of a sign

Description generated with very high confidence

tail(actuals\_preds)

A close up of a sign

Description generated with very high confidence

min\_max\_accuracy <- mean (apply(actuals\_preds, 1, min) / apply(actuals\_preds, 1, max))

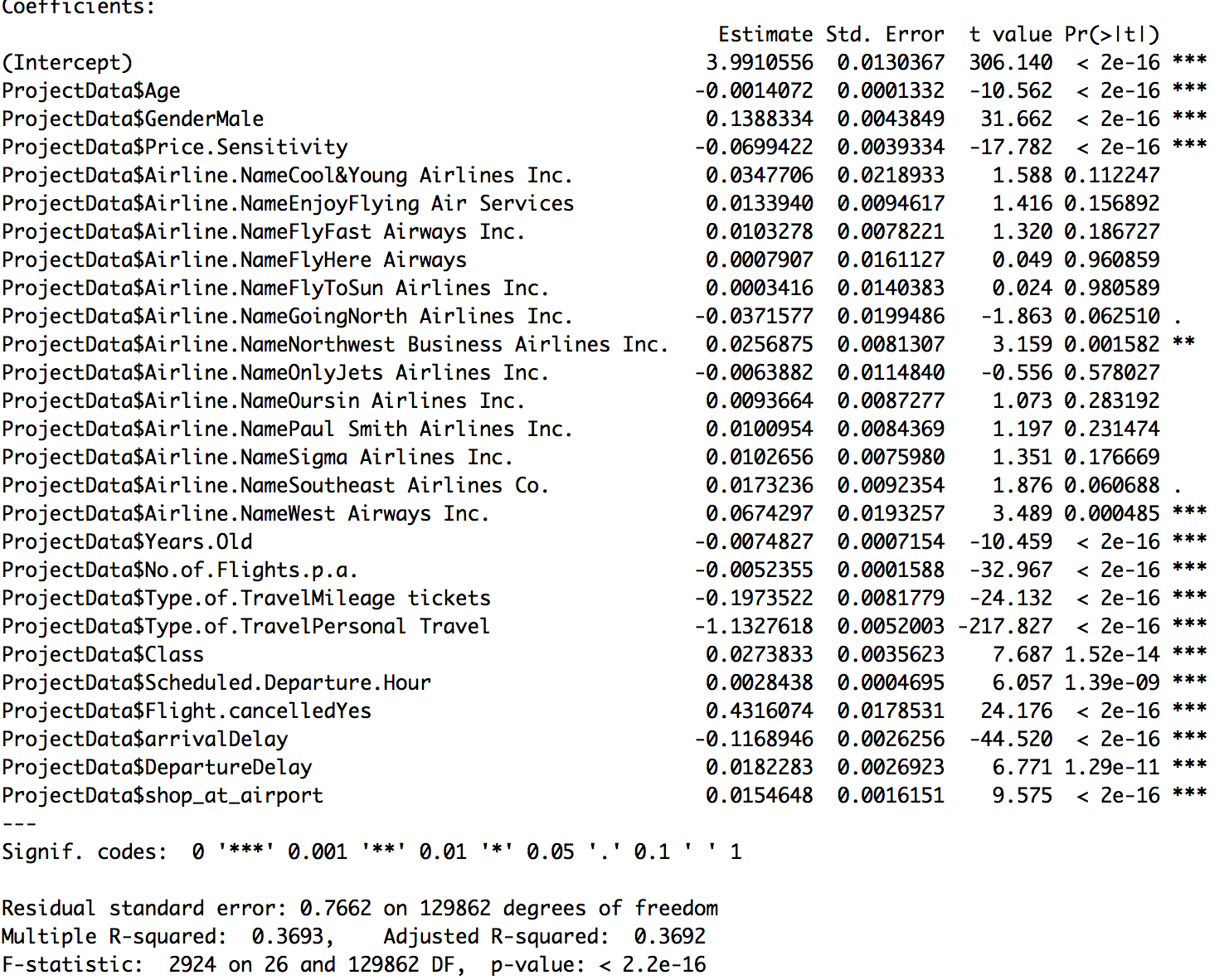
min\_max\_accuracy

A black and blue text

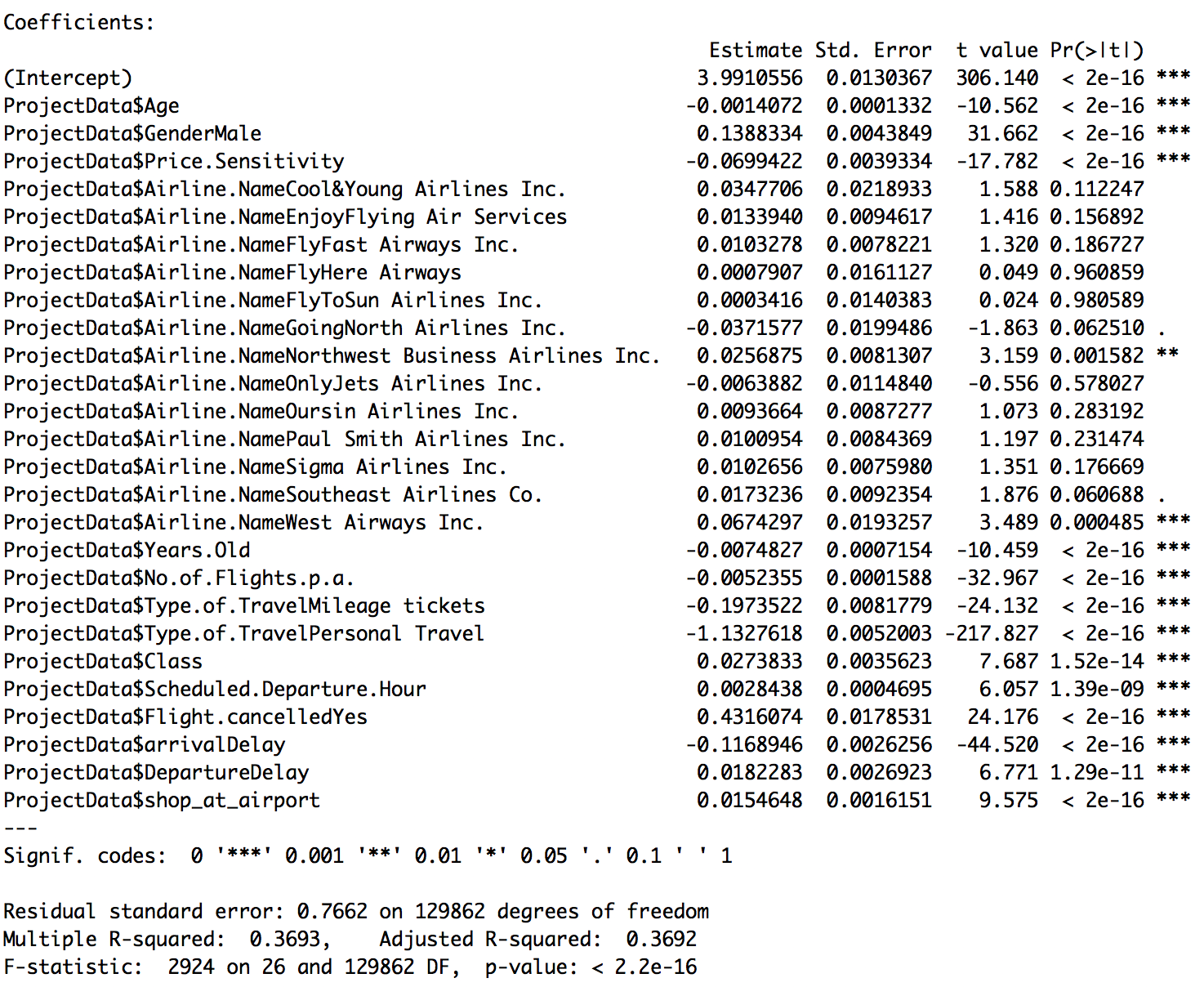
Description generated with high confidence

# Linear Regression

At the first, we need to find out significant variables from massive size of data. it’s kind of challenge for us to look deeper inside each variable, so we start with running a full model, and then find out significant variables, after that we can fit them into our reduce model.

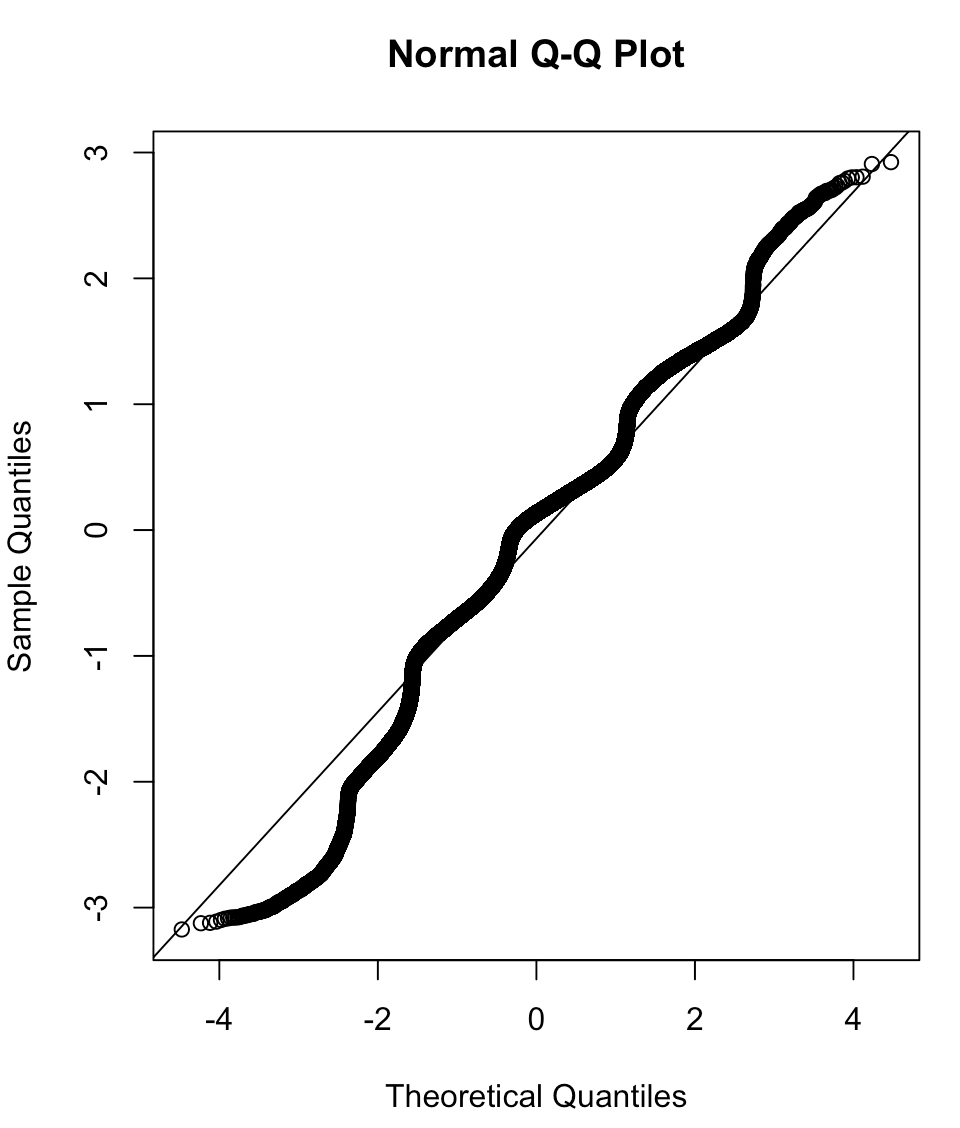


Base on the result above, we can find survey significant components are: Age, Gendemale, Pricesensitivity, No.of.Flights.p.a., Type.of.Travelmileage tickets, Type.of.Travelpersonal Travel, Class, Scheduleddeparturehour, Flightcancelled, arrivaldelay, Departuredelay, shopatairport. Because their p-value<0.05, so we can say there are significant with 95% confident.



Reduce model: according to the result of full model, we collect all the significant variables, then refit them into our reduce model, here is the result,

Now we can find all significant variables, according to our r square, there is no significant difference between 0.4028 and 0.3692, we could possibly accept this model.



Based on our residual plot above, although our reduce model is not perfectly fit, but the residual is close to normal distribution, so if prove our reduce model is acceptable.

*ProjectData <-read.csv("Satisfaction Survey.CSV")*

*Model <- lm(as.numeric(Satisfaction) ~.,data = ProjectData)*

*summary(Model)*

*ProjectData$AirlineStatus.code<- factor(ProjectData$AirlineStatus.code)*

*Model2<- lm(as.numeric(Satisfaction) ~ ProjectData$Age +ProjectData$Gender +ProjectData$Price.Sensitivity +ProjectData$Years.Old +ProjectData$No.of.Flights.p.a.+ProjectData$Type.of.Travel+ProjectData$Type.of.Travel+ProjectData$Class+ProjectData$Scheduled.Departure.Hour+ProjectData$Flight.cancelled+ProjectData$arrivalDelay+ProjectData$DepartureDelay+ProjectData$shop\_at\_airport,data = ProjectData)*

*summary(Model2)*

*qqnorm(Model2$residuals)*

*qqline(Model2$residuals)*

# Association Rules

1. Association Rules only work the data that is not numerical, so we had to convert numerical variable that are not converted during Pre-Processing Phase.

*Create\_AirplainSatisfaction <- function(vec){*

*vBuckets <- replicate(length(vec), "high")*

*vBuckets[vec <= 3 & vec > 2] <- "medium"*

*vBuckets[vec <= 2] <- "low"*

*return(vBuckets)*

*}*

*age <- function(vec){*

*vBuckets <- replicate(length(vec), "Adolescence")*

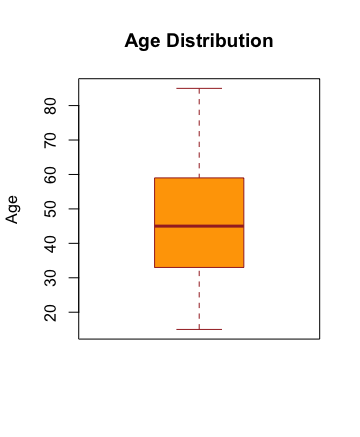
*vBuckets[vec <= 35 & vec > 20] <- "Young Adult"*

*vBuckets[vec <= 55 & vec > 35] <- "Senior Adult"*

*vBuckets[vec > 55] <- "old"*

*return(vBuckets)*

*}*



*yrol <- function(vec){*

*vBuckets <- replicate(length(vec), "new")*

*vBuckets[vec > 3] <- "old"*

*return(vBuckets)*

*}*

*survey$Satisfaction <- as.factor(Create\_AirplainSatisfaction(survey$Satisfaction))*

*survey$Age <- as.factor(age(survey$Age))*

*survey$Years.Old <- as.factor(yrol(survey$Years.Old))*

*survey$Price.Sensitivity <- as.factor(Create\_AirplainSatisfaction(survey$Price.Sensitivity))*

1. Importing the packages used for finding Association Rules in R.

library(arules)

library(arulesViz)

1. First Finding association when satisfaction is Low.

* Support = 0.01
* Confidence = 0.5
* Lyft = 4.6

*ruleset <- apriori(survey,parameter = list(support=0.01, confidence = 0.5),*

*appearance = list(default="lhs",rhs=("Satisfaction=low")))*

*goodrules <- ruleset[quality(ruleset)$lift > 4.65]*

*inspect(goodrules*)

1. Rules for 3rd Step

**LHS RHS Support Confidence Lift Count**

[1] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

DepartureDelay=2 hours} => {Satisfaction=low} 0.01018562 0.9416370 4.600477 1323

[2] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

arrivalDelay=2 hours} => {Satisfaction=low} 0.01057056 0.9495159 4.638971 1373

[3] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

DepartureDelay=1 hours} => {Satisfaction=low} 0.01657569 0.9463736 4.623619 2153

[4] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

arrivalDelay=1 hours} => {Satisfaction=low} 0.01629853 0.9536036 4.658941 2117

[5] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

Flight.cancelled=No,

DepartureDelay=2 hours} => {Satisfaction=low} 0.01018562 0.9456755 4.620208 1323

[6] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

Flight.cancelled=No,

arrivalDelay=2 hours} => {Satisfaction=low} 0.01057056 0.9495159 4.638971 1373

[7] {Airline.Status=Blue,

Price.Sensitivity=low,

Type.of.Travel=Personal Travel,

arrivalDelay=2 hours} => {Satisfaction=low} 0.01029340 0.9482270 4.632673 1337

[8] {Airline.Status=Blue,

Gender=Female,

Type.of.Travel=Personal Travel,

DepartureDelay=1 hours} => {Satisfaction=low} 0.01148673 0.9570237 4.675651 1492

[9] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

DepartureDelay=1 hours,

shop\_at\_airport=No shopping} => {Satisfaction=low} 0.01034730 0.9464789 4.624133 1344

[10] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

Class=Eco,

DepartureDelay=1 hours} => {Satisfaction=low} 0.01348074 0.9459751 4.621672 1751

[11] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

Flight.cancelled=No,

DepartureDelay=1 hours} => {Satisfaction=low} 0.01654490 0.9475309 4.629272 2149

[12] {Airline.Status=Blue,

Price.Sensitivity=low,

Type.of.Travel=Personal Travel,

DepartureDelay=1 hours} => {Satisfaction=low} 0.01615225 0.9454709 4.619208 2098

[13] {Airline.Status=Blue,

Gender=Female,

Type.of.Travel=Personal Travel,

arrivalDelay=1 hours} => {Satisfaction=low} 0.01122497 0.9655629 4.717370 1458

[14] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

arrivalDelay=1 hours,

shop\_at\_airport=No shopping} => {Satisfaction=low} 0.01013173 0.9570909 4.675979 1316

[15] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

Class=Eco,

arrivalDelay=1 hours} => {Satisfaction=low} 0.01320358 0.9538376 4.660085 1715

[16] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

Flight.cancelled=No,

arrivalDelay=1 hours} => {Satisfaction=low} 0.01629853 0.9536036 4.658941 2117

[17] {Airline.Status=Blue,

Price.Sensitivity=low,

Type.of.Travel=Personal Travel,

arrivalDelay=1 hours} => {Satisfaction=low} 0.01586739 0.9528433 4.655227 2061

[18] {Airline.Status=Blue,

Age=old,

Type.of.Travel=Personal Travel,

arrivalDelay=half hours} => {Satisfaction=low} 0.02632248 0.9431724 4.607979 3419

[19] {Airline.Status=Blue,

Price.Sensitivity=low,

Type.of.Travel=Personal Travel,

Flight.cancelled=No,

arrivalDelay=2 hours} => {Satisfaction=low} 0.01029340 0.9482270 4.632673 1337

[20] {Airline.Status=Blue,

Gender=Female,

Type.of.Travel=Personal Travel,

Flight.cancelled=No,

DepartureDelay=1 hours} => {Satisfaction=low} 0.01145594 0.9581455 4.681132 1488

[21] {Airline.Status=Blue,

Gender=Female,

Price.Sensitivity=low,

Type.of.Travel=Personal Travel,

DepartureDelay=1 hours} => {Satisfaction=low} 0.01123267 0.9560944 4.671110 1459

[22] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

Flight.cancelled=No,

DepartureDelay=1 hours,

shop\_at\_airport=No shopping} => {Satisfaction=low} 0.01033960 0.9477770 4.630475 1343

[23] {Airline.Status=Blue,

Price.Sensitivity=low,

Type.of.Travel=Personal Travel,

DepartureDelay=1 hours,

shop\_at\_airport=No shopping} => {Satisfaction=low} 0.01000855 0.9454545 4.619128 1300

[24] {Airline.Status=Blue,

Type.of.Travel=Personal Travel,

Class=Eco,

Flight.cancelled=No,

DepartureDelay=1 hours} => {Satisfaction=low} 0.01345764 0.9469122 4.626250 1748

[25] {Airline.Status=Blue,

Price.Sensitivity=low,

Type.of.Travel=Personal Travel,

Class=Eco,

DepartureDelay=1 hours} => {Satisfaction=low} 0.01316509 0.9452736 4.618244 1710

1. Most common things that are associated with lower ratings are

* Type of travel = Personal
* Price Sensitivity = Low
* Age = Old
* Shop at Airport = No Shopping
* Eating and Drinking = No
* Flight Cancelled = No
* Year Old = Old
* Class = Eco
* Arrival Delays = Half hour or one hour
* Departure Delays = Half hour or one hour
* Airline Status = Blue

1. Finding association when satisfaction is High.

* Support = 0.006
* Confidence = 0.4
* Lyft = 1.95

ruleset <- apriori(survey,parameter = list(support=0.006, confidence = 0.4),

appearance = list(default="lhs",rhs=("Satisfaction=high")))

goodrules <- ruleset[quality(ruleset)$lift > 1.95]

inspect(goodrules)

1. Rules for step 6th

LHS RHS support confidence lift count

[1] {Airline.Status=Silver,

Type.of.Travel=Mileage tickets} => {Satisfaction=high} 0.011871675 1 1.958608 1542

[2] {Airline.Status=Silver,

Age=Senior Adult,

Type.of.Travel=Mileage tickets} => {Satisfaction=high} 0.006066719 1 1.958608 788

[3] {Airline.Status=Silver,

Type.of.Travel=Mileage tickets,

arrivalDelay=No Delay} => {Satisfaction=high} 0.006351577 1 1.958608 825

[4] {Airline.Status=Silver,

Type.of.Travel=Mileage tickets,

DepartureDelay=No Delay} => {Satisfaction=high} 0.006382373 1 1.958608 829

[5] {Airline.Status=Silver,

Gender=Female,

Type.of.Travel=Mileage tickets} => {Satisfaction=high} 0.006698027 1 1.958608 870

[6] {Airline.Status=Silver,

Years.Old=old,

Type.of.Travel=Mileage tickets} => {Satisfaction=high} 0.006420867 1 1.958608 834

[7] {Airline.Status=Silver,

Type.of.Travel=Mileage tickets,

Class=Eco} => {Satisfaction=high} 0.009846869 1 1.958608 1279

[8] {Airline.Status=Silver,

Type.of.Travel=Mileage tickets,

Flight.cancelled=No} => {Satisfaction=high} 0.011602214 1 1.958608 1507

[9] {Airline.Status=Silver,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets} => {Satisfaction=high} 0.011771590 1 1.958608 1529

[10] {Airline.Status=Silver,

Age=Senior Adult,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets} => {Satisfaction=high} 0.006028224 1 1.958608 783

[11] {Airline.Status=Silver,

Type.of.Travel=Mileage tickets,

Flight.cancelled=No,

arrivalDelay=No Delay} => {Satisfaction=high} 0.006351577 1 1.958608 825

[12] {Airline.Status=Silver,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets,

arrivalDelay=No Delay} => {Satisfaction=high} 0.006305384 1 1.958608 819

[13] {Airline.Status=Silver,

Type.of.Travel=Mileage tickets,

Flight.cancelled=No,

DepartureDelay=No Delay} => {Satisfaction=high} 0.006382373 1 1.958608 829

[14] {Airline.Status=Silver,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets,

DepartureDelay=No Delay} => {Satisfaction=high} 0.006336179 1 1.958608 823

[15] {Airline.Status=Silver,

Gender=Female,

Type.of.Travel=Mileage tickets,

Flight.cancelled=No} => {Satisfaction=high} 0.006544049 1 1.958608 850

[16] {Airline.Status=Silver,

Gender=Female,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets} => {Satisfaction=high} 0.006644135 1 1.958608 863

[17] {Airline.Status=Silver,

Years.Old=old,

Type.of.Travel=Mileage tickets,

Flight.cancelled=No} => {Satisfaction=high} 0.006266889 1 1.958608 814

[18] {Airline.Status=Silver,

Price.Sensitivity=low,

Years.Old=old,

Type.of.Travel=Mileage tickets} => {Satisfaction=high} 0.006359276 1 1.958608 826

[19] {Airline.Status=Silver,

Type.of.Travel=Mileage tickets,

Class=Eco,

Flight.cancelled=No} => {Satisfaction=high} 0.009592806 1 1.958608 1246

[20] {Airline.Status=Silver,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets,

Class=Eco} => {Satisfaction=high} 0.009769880 1 1.958608 1269

[21] {Airline.Status=Silver,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets,

Flight.cancelled=No} => {Satisfaction=high} 0.011502129 1 1.958608 1494

[22] {Airline.Status=Silver,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets,

Flight.cancelled=No,

arrivalDelay=No Delay} => {Satisfaction=high} 0.006305384 1 1.958608 819

[23] {Airline.Status=Silver,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets,

Flight.cancelled=No,

DepartureDelay=No Delay} => {Satisfaction=high} 0.006336179 1 1.958608 823

[24] {Airline.Status=Silver,

Gender=Female,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets,

Flight.cancelled=No} => {Satisfaction=high} 0.006490157 1 1.958608 843

[25] {Airline.Status=Silver,

Price.Sensitivity=low,

Years.Old=old,

Type.of.Travel=Mileage tickets,

Flight.cancelled=No} => {Satisfaction=high} 0.006205298 1 1.958608 806

[26] {Airline.Status=Silver,

Price.Sensitivity=low,

Type.of.Travel=Mileage tickets,

Class=Eco,

Flight.cancelled=No} => {Satisfaction=high} 0.009515817 1 1.958608 1236

1. Most common things that are associated with Higher ratings are

* Type of Travel = Mileage Tickets
* Airline Status = Silver
* Departure Delay = No Delay
* Price Sensitivity = Low
* Flight Cancelled = No

# Decision Tree

* Definition: A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.
* Before doing the decision tree analysis, we run the multiple regression model

By the p value, we delete

* No.of.other.Loyalty.Cards,
* Airline.Name,
* Flight.Distance,
* weekday

are not significant so we delete those.

AirplainSatisfaction$No..of.other.Loyalty.Cards <- NULL

AirplainSatisfaction$Flight.Distance <- NULL

AirplainSatisfaction$Airline.Name <- NULL

AirplainSatisfaction$weekday <- NULL

AirplainSatisfaction$X <- NULL

* In order to fit decision tree model, we transfer the column “Satisfaction” to category. We transfer to three categories. High, medium, and low.

Create\_AirplainSatisfaction <- function(vec){

vBuckets <- replicate(length(vec), "high")

vBuckets[vec <= 3 & vec > 2] <- "medium"

vBuckets[vec <= 2] <- "low"

return(vBuckets)

}

AirplainSatisfaction$Evaluation <- Create\_AirplainSatisfaction(AirplainSatisfaction$Satisfaction)

AirplainSatisfaction$Satisfaction <- NULL

str(AirplainSatisfaction)

* Here is the decision tree. In order to compare the subset and the whole dataset, we also split dataset to three parts. From the photos below, we could see the results of subsets are very similar to the whole dataset.

library(rpart)

library(rpart.plot)

fit <- rpart(Evaluation~., data = AirplainSatisfaction, method = 'class', cp=0.005)

rpart.plot(fit)

a <- nrow(AirplainSatisfaction)

a1 <- a/3

df1 <- AirplainSatisfaction[1:a1,]

a2 <- a/3 \* 2

df2 <- AirplainSatisfaction[a1:a2,]

a3 <- a/3 \* 3

df3 <- AirplainSatisfaction[a2:a3,]

#df1

fit <- rpart(Evaluation~., data = df1, method = 'class', cp=0.005)

rpart.plot(fit)

#df2

fit <- rpart(Evaluation~., data = df2, method = 'class', cp=0.005)

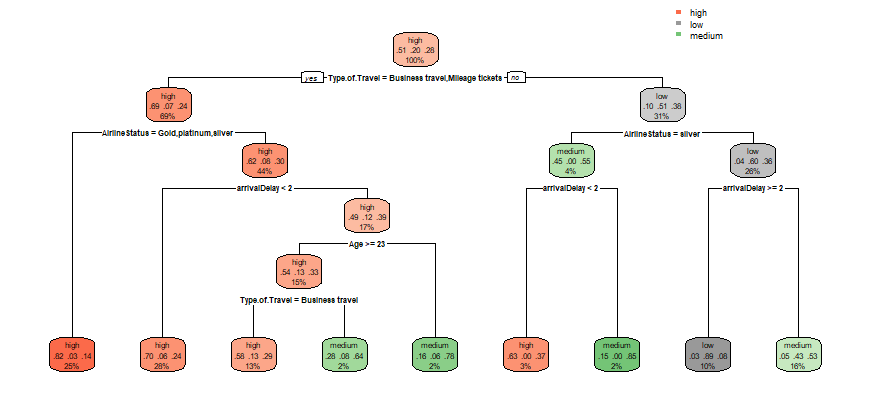
rpart.plot(fit)

#df3

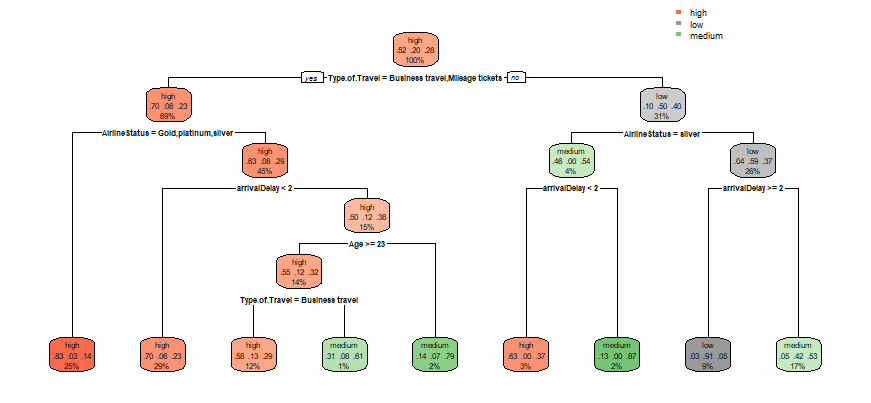
fit <- rpart(Evaluation~., data = df3, method = 'class', cp=0.005)

rpart.plot(fit)

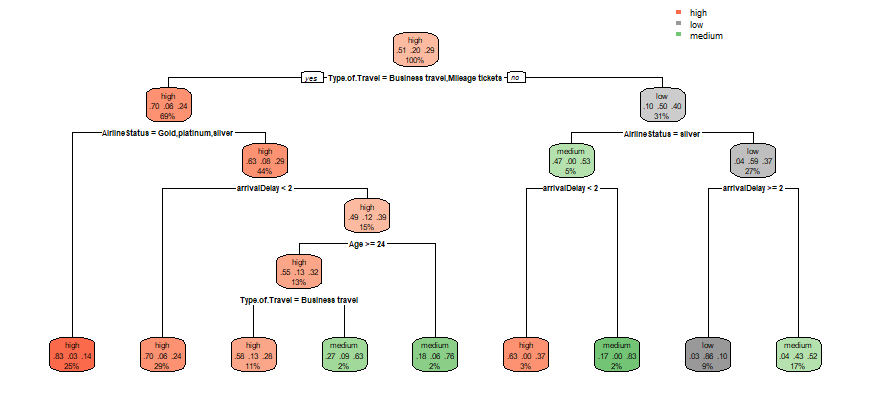
* All Data:



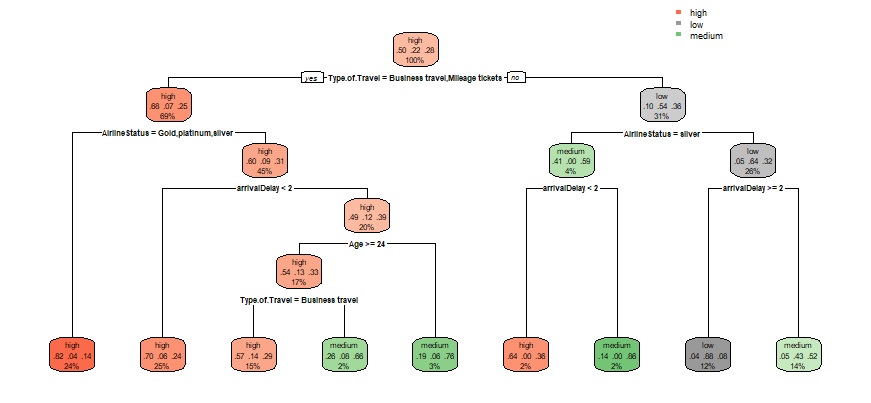
* Split\_df1



* Split\_df2



* Split\_df3



Analysis of the Results:

* From the result, we could see 51% of customers have high satisfaction.
* The first condition is “Type of Travel”. If the customers are Business travel and Mileage tickets, there are 69% of customers with high satisfaction probability of 69%. If the customers are not Business travel and Mileage tickets, there are only 31% of customers have high satisfaction.
* Important factors that affect the distribution of satisfied customers with not so satisfied customers.
* Type of Travel
* Airline Status
* Arrival Delay
* Age
* Traits of a customer that is highly satisfied.
* Type of travel = Business and mileage
* Arrival Delay = less than 2 hours
* Age = greater than 24 (that is adult or old)
* Traits of a customer that is not satisfied
* Type of travel = Personal
* Arrival Delay = greater than 2 hours

# Support Vector Machine

1. Since we had a large dataset and analysis was done based on the whole dataset, we decided to use a random sample of size 30,000 for SVM.

surveyData<-read.csv("surveyDataV7.csv")

surveyData<- data.frame(surveyData) #read the data

View(surveyData)

surveySampleIndex <- sample(1:dim(surveyData)[1],size=30000) #created a sample for surveyData of size 30000

newsurveyData <- surveyData[surveySampleIndex,]

View(newsurveyData)

1. We cleaned the data and only kept attributes which are important for SVM.

newsurveyData<-newsurveyData[,-1]

newsurveyData<-newsurveyData[,-8] #removing loayalty cards

newsurveyData <-newsurveyData[,-12] #removing flight distance

str(newsurveyData)

1. Since we required Satisfaction as factors for svm, we converted it into high, medium and low values.

Create\_AirplainSatisfaction <- function(vec){

vBuckets <- replicate(length(vec), "medium")

vBuckets[vec > 3] <- "high"

vBuckets[vec <= 2] <- "low"

return(vBuckets)

}

newsurveyData$Satisfaction <- as.factor(Create\_AirplainSatisfaction(newsurveyData$Satisfaction))

1. We performed KSVM on the sample with 2/3rd of the sample data for training and 1/3rd for testing. We considered Cost =10 and 10-fold cross validation.

cutPoint <- floor(2\*dim(newsurveyData)[1]/3)

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

summary(randIndex)

train <- newsurveyData[randIndex[1:cutPoint],]

test <- newsurveyData[randIndex[(cutPoint+1):dim(newsurveyData)[1]],]

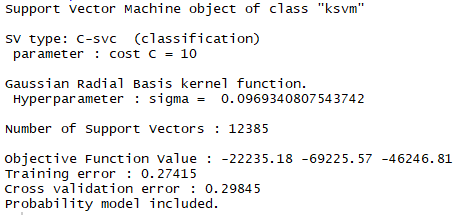
dim(train)#20000

dim(test) #10000

library(kernlab)

svmOutput <- ksvm(Satisfaction ~Type.of.Travel + Price.Sensitivity + Age + shop\_at\_airport + Flight.cancelled + Years.Old + Class + arrivalDelay + DepartureDelay + AirlineStatus, data= train, kernel= "rbfdot", kpar = "automatic", C = 10, cross = 10, prob.model = TRUE)

svmOutput



svmPred <- predict(svmOutput, test)

svmPred <- as.numeric(svmPred)

1. We obtained the accuracy of the model by calculating the Error Rate

ctable1<- data.frame(test$Satisfaction,svmPred)

tableC<-table(ctable1)

t1<-tableC[1,1]+tableC[2,2]+tableC[3,3]

t2<-tableC[1,2]+tableC[1,3]+tableC[2,1]+tableC[2,3]+tableC[3,1]+tableC[3,2]

accuracy <- (t1/(t1+t2)) \* 100

# Accuracy is 70.57

**Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
| classes | High | Low | Medium |
| high | **4806** | 54 | 227 |
| low | 397 | **992** | 666 |
| Medium | 1427 | 172 | **1259** |

* This High accuracy in Support Vector Machine shows that the variables we considered were accurate and the Data Preprocessing was appropriate.

# Recommended Significant Variables for Future Models

|  |  |  |
| --- | --- | --- |
| **Variables** | **Type** | **Sample Values** |
| Airline Status | Factors => 4 Levels | Blue, Gold, silver, Platinum |
| Age | Factors => 4 Levels | Adolescence, Young Adult, old, Senior Adult |
| Gender | Factors => 2 Levels | Male, Female |
| Price Sensitivity | int | Range 0 to 5 |
| Years old (age of plane) | int | Range 0 to 8 |
| Class | Factors => 3 Levels | Business, eco, Eco Plus |
| Type of Travel | Factors => 3 Levels | Business, Personal, Mileage tickets |
| Arrival Delay | Factors => 7 Levels | 1 hours, 2 hours, cancelled, less than 5 minutes, cancelled, half hours |
| Departure Delay | Factors => 7 Levels | 1 hours, 2 hours, cancelled, less than 5 minutes, cancelled, half hours |
| Shopping at Airport | Factors => 5 Levels | No Shopping, less than 20$, more than 80$,.. |

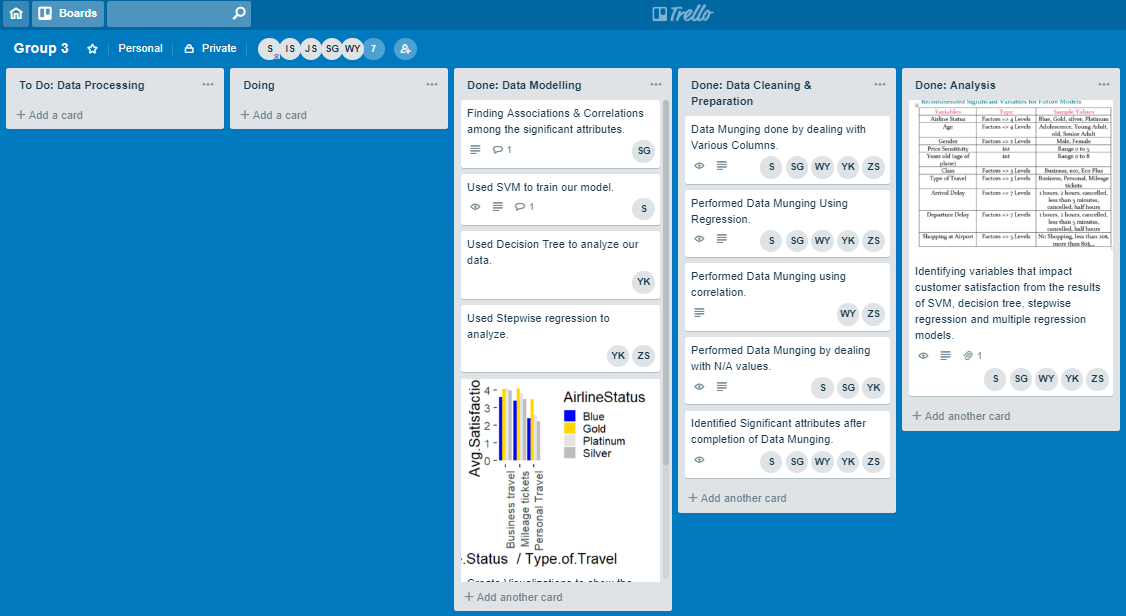
Insights and Recommendation for Airlines Form these Models and Association Rules

* Natural Tendencies that affect Satisfaction Rating But cannot be Changed
* Highly satisfied Customers tend to be in
* Age: Between 25 to 55
* Gender: Male
* Years old: Plane not older than 3 years
* Airline Status: Silver, Gold, Platinum
* Not so satisfied Customers tend to in
* Age: Young Or older than 55
* Gender: Female
* Years old: Travelling in Older planes
* Airline Status: Blue
* No of flights p.a: less than 5
* Things That can be Worked on to improve Satisfaction Rating
* People with High Satisfaction
* Arrival delay: No Delay
* Departure Delay: No Delay
* Price Sensitivity: 0 to 3
* Shopping at Airport: They do Shop at Airport
* Type of Travel: Mileage Tickets
* Things to work on as these cause Lower Ratings
* Arrival Delay: More than two Hours
* Departure Delay: More than two Hours
* Flight Cancelled: Yes
* Type of Travel: Personal
* Price Sensitivity: 3 to 5
* Shopping and Eating: No Shopping & No Eating

# Actionable Insights

* Provide reward points to people with Mileage Tickets for retaining the satisfied people.
* The travel delay for type of traveler: personal should be reduced to less than two hours.
* If the flight needs to be canceled, the airlines could provide discounts on the travel tickets to the travelers. The discounts will help improve customer satisfaction for people impacted by high price sensitivity.
* Provide discount coupons on shopping at Airport for travelers.
* Take travel feedback from people above age 60.

# Trello and Work Division



Details and Works Distribution

* Work Distribution
* Shloak Gupta: Data Preprocessing, Association Rules, Insights and Conclusions
* Yu Yu Ko: Decision Tree, Stepwise Regression, Data Visualizations
* Satyen Amorkar: Support Vector Machine, Data Visualizations, Kanban Master
* Zhenyu Sao: Linear Regression
* Wenjin Yao: Data Visualization, Data Processing
* Walk through the Process
* Started working together on the data and we kept distributing work among us.
* Everyone was able to complete their work before next meeting schedules.
* We complied the work in meeting and discussed our work among ourselves.
* In the end all of the models got similar results which shows the coordination and validates our team work and the same thought process.