ANALYTICS ON AVIATION DATA: TURNING DATA INTO INSIGHTS

AIT 582 (Final Project)

ANIRUDH MYAKALA

1 TABLE OF CONTENTS

2 INTRODUCTION

An airline customer database with various attributes wants to identify the factors that are helpful to understand why some customers are flying and why others are canceling. I as a Data Scientist used the metadata along with other attributes in creating a classification/prediction model which will be able to predict if a customer will fly or not given his attributes and thus made a list of recommendations that can be given to the advertising team, such as customer demographic specifics packages to attract more customers.

2.2 Tools

1.

R language was implemented in RStudio IDE to collect, pre-process and clean the data apart from performing some basic statically analysis.

2.

Tableau was used to perform meta data exploration. And also, to derive some visual insights.

3.

Weka was used to prepare the data for data mining algorithm and develop classification models.

3 MILESTONE-1 (Data Acquisition and Conversion)

3.2 Programmatically downloading the project data file

The data file in JSON format was downloaded from the below link using R
http://ist.gmu.edu/~hpurohit/courses/ait582-proj-data-spring16.json

3.3 JSON to CSV

Since for easier data manipulation and interpretation JSON file is not that great I converted the JSON into CSV format

3.4 Output Fields

	A	B	C	D	E	F	G
1	FARE	DESCRIPTION	SUCCESS	SEATCLAS	GUESTS	CUSTOMERID	
2	7.25	Braund, Mr. Owen Harris;22	0	3	1	1	
3	71.2833	Cumings, Mrs. John Bradley (Florence Briggs Thayer);38	1	1	1	2	
4	7.925	Heikkinen, Miss. Laina;26	1	3	0	3	
5	53.1	Futrelle, Mrs. Jacques Heath (Lily May Peel);35	1	1	1	4	
6	8.05	Allen, Mr. William Henry;35	0	3	0	5	
7	8.4583	Moran, Mr. James;	0	3	0	6	
8	51.8625	McCarthy, Mr. Timothy J;54	0	1	0	7	
9	21.075	Palsson, Master. Gosta Leonard;2	0	3	3	8	
10	11.1333	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg);27	1	3	0	9	
11	30.0708	Nasser, Mrs. Nicholas (Adele Achem);14	1	2	1	10	
12	16.7	Sandstrom, Miss. Marguerite Rut;4	1	3	1	11	
13	26.55	Bonnell, Miss. Elizabeth;58	1	1	0	12	
14	8.05	Saundercock, Mr. William Henry;20	0	3	0	13	
15	31.275	Andersson, Mr. Anders Johan;39	0	3	1	14	
16	7.8542	Vestrom, Miss. Hulda Amanda Adolfina;14	0	3	0	15	
17	16	Hewlett, Mrs. (Mary D Kingcome) ;55	1	2	0	16	
18	29.125	Rice, Master. Eugene;2	0	3	4	17	
19	13	Williams, Mr. Charles Eugene;	1	2	0	18	
20	18	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele);31	0	3	1	19	
21	7.225	Masselmani, Mrs. Fatima;	1	3	0	20	
22	26	Fynney, Mr. Joseph J;35	0	2	0	21	
23	13	Beesley, Mr. Lawrence;34	1	2	0	22	

There are 6 output fields namely
1.Fare
2.Description
3.Success
4.Seatclass
5.Guests
6.CustomerId

4 MILESTONE-2 (Metadata Extraction and Imputation)

4.2 Extracting Metadata

Clearly looking at the Description column we can see that there is potential meta data hidden that could be the extracted and used as attributes.

Using R I extracted metadata in description to age and title based on 4 title factors i.e. Mr., Mrs., Master., Miss. The entries around 17 with other title were removed as they have little significance.

	FARE	SUCCESS	SEATCLASS	GUESTS	CUSTOMERID	GENDER	AGE
56	110.8833	1	1	0	307	Miss.	NA
57	110.8833	1	1	0	551	Mr.	17.0
58	110.8833	1	1	1	582	Mrs.	39.0
59	113.275	0	1	0	660	Mr.	58.0
60	113.275	1	1	1	216	Miss.	31.0
61	113.275	1	1	1	394	Miss.	23.0
62	12	1	2	0	390	Miss.	17.0
63	12.275	0	2	0	240	Mr.	33.0
64	12.2875	1	3	0	480	Miss.	2.0
65	12.35	1	2	0	304	Miss.	NA
66	12.35	1	2	0	323	Miss.	30.0
67	12.475	1	3	0	752	Master.	6.0

4.3 Missing Values Imputation

Clearly looking at the age field we can see that AGE has many missing values. And these values must be imputed by some technique. But after clearly knowing the separation in AGE based on TITLE it would be bad analytics to impute data as it is.

So, for the Imputation I splitted the data into 4 tables to apply imputation techniques.

1.Mr

	FARE	SUCCESS	SEATCLASS	GUESTS	CUSTOMERID	GENDER	AGE
1	7.25	0	3	1	1	Mr.	22.0
2	8.05	0	3	0	5	Mr.	35.0
3	8.4583	0	3	0	6	Mr.	NA
4	51.8625	0	1	0	7	Mr.	54.0
5	8.05	0	3	0	13	Mr.	20.0
6	31.275	0	3	1	14	Mr.	39.0
7	13	1	2	0	18	Mr.	NA
8	26	0	2	0	21	Mr.	35.0
9	13	1	2	0	22	Mr.	34.0
10	35.5	1	1	0	24	Mr.	28.0
11	7.225	0	3	0	27	Mr.	NA
12	263	0	1	3	28	Mr.	19.0

2.Mrs

	FARE	SUCCESS	SEATCLASS	GUESTS	CUSTOMERID	GENDER	AGE
1	71.2833	1	1	1	2	Mrs.	38
2	53.1	1	1	1	4	Mrs.	35
3	11.1333	1	3	0	9	Mrs.	27
4	30.0708	1	2	1	10	Mrs.	14
5	16	1	2	0	16	Mrs.	55
6	18	0	3	1	19	Mrs.	31
7	7.225	1	3	0	20	Mrs.	NA
8	31.3875	1	3	1	26	Mrs.	38
9	146.5208	1	1	1	32	Mrs.	NA
10	9.475	0	3	1	41	Mrs.	40
11	21	0	2	1	42	Mrs.	27
12	17.8	0	3	1	50	Mrs.	18

3.Master

	FARE	SUCCESS	SEATCLASS	GUESTS	CUSTOMERID	GENDER	AGE
1	21.075	0	3	3	8	Master.	2.00
2	29.125	0	3	4	17	Master.	2.00
3	39.6875	0	3	4	51	Master.	7.00
4	46.9	0	3	5	60	Master.	11.00
5	27.9	0	3	3	64	Master.	4.00
6	15.2458	1	3	1	66	Master.	*NA*
7	29	1	2	0	79	Master.	0.83
8	11.2417	1	3	1	126	Master.	12.00
9	69.55	0	3	8	160	Master.	*NA*
10	39.6875	0	3	4	165	Master.	1.00
11	20.525	1	3	0	166	Master.	9.00
12	29.125	0	3	4	172	Master.	4.00

4.Miss

	FARE	SUCCESS	SEATCLASS	GUESTS	CUSTOMERID	GENDER	AGE
1	7.925	1	3	0	3	Miss.	26.00
2	16.7	1	3	1	11	Miss.	4.00
3	26.55	1	1	0	12	Miss.	58.00
4	7.8542	0	3	0	15	Miss.	14.00
5	8.0292	1	3	0	23	Miss.	15.00
6	21.075	0	3	3	25	Miss.	8.00
7	7.8792	1	3	0	29	Miss.	*NA*
8	7.75	1	3	0	33	Miss.	*NA*
9	18	0	3	2	39	Miss.	18.00
10	11.2417	1	3	1	40	Miss.	14.00
11	41.5792	1	2	1	44	Miss.	3.00
12	7.8792	1	3	0	45	Miss.	19.00

Now, predictive mean matching(PMM) using mice (Multivariate Imputation by Chained Equations) package was performed on each category separately to generate unique value for each missing value.

After the missing values were imputed the labels for title were change.
For Mr., Master- MALE was assigned and for Mrs., Miss -FEMALE was assigned.
Here is what the final clean dataset looks like after removing insignificant values and imputing missing age values.

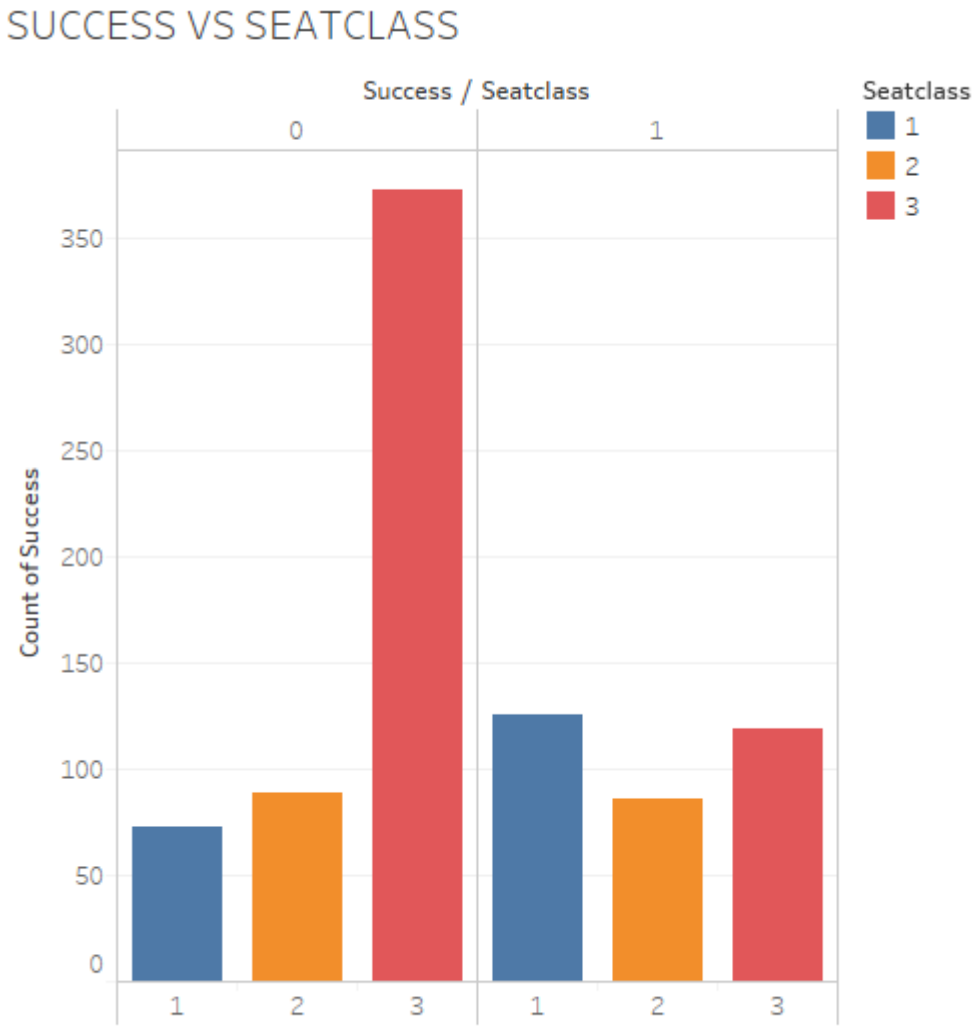
	FARE	SUCCESS	SEATCLASS	GUESTS	CUSTOMERID	GENDER	AGE
24	10.5	0	2	0	440	MALE	31.0
25	10.5	0	2	0	620	MALE	26.0
26	10.5	0	2	0	673	MALE	70.0
27	10.5	0	2	0	71	MALE	32.0
28	10.5	0	2	0	773	FEMALE	57.0
29	10.5	0	2	0	813	MALE	35.0
30	10.5	0	2	0	842	MALE	16.0
31	10.5	0	2	0	884	MALE	28.0
32	10.5	1	2	0	227	MALE	19.0
33	10.5	1	2	0	459	FEMALE	50.0
34	10.5	1	2	0	517	FEMALE	34.0
35	10.5	1	2	0	527	FEMALE	50.0

Now data is perfect to perform any sough of analytics.

5 MILESTONE-3 (Metadata Exploration)

Meta data exploration was done on how success was varying with all the other attributes.

1.SUCCESS VS SEATCLASS



Count of Success for each Seatclass broken down by Success. Color shows details about Seatclass.

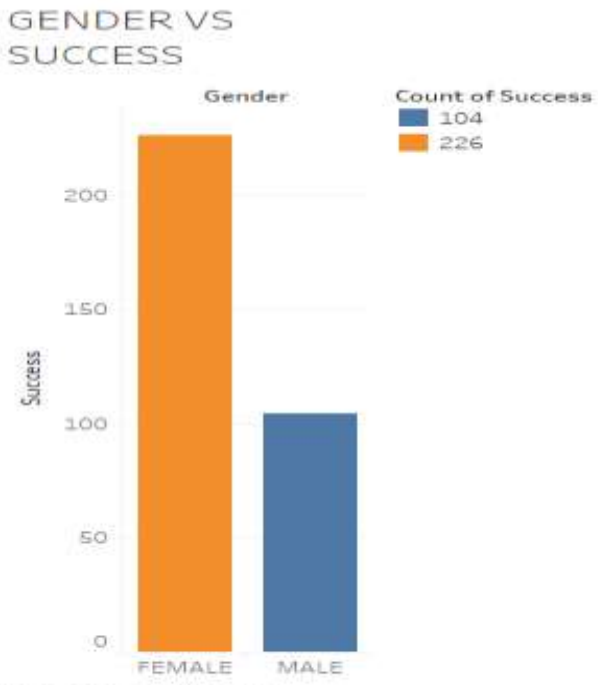
- Seat class 1 is highly preferred
- Seat class 3 is least preferred

2.AGE VS SUCCESS

AGE. Color shows count of Success. Size shows count of Success. The marks are labeled by AGE. The data is filtered on Success, which keeps 1.

- People around the age of 24 are frequent fliers
- The age group with highest success is around 25-35.

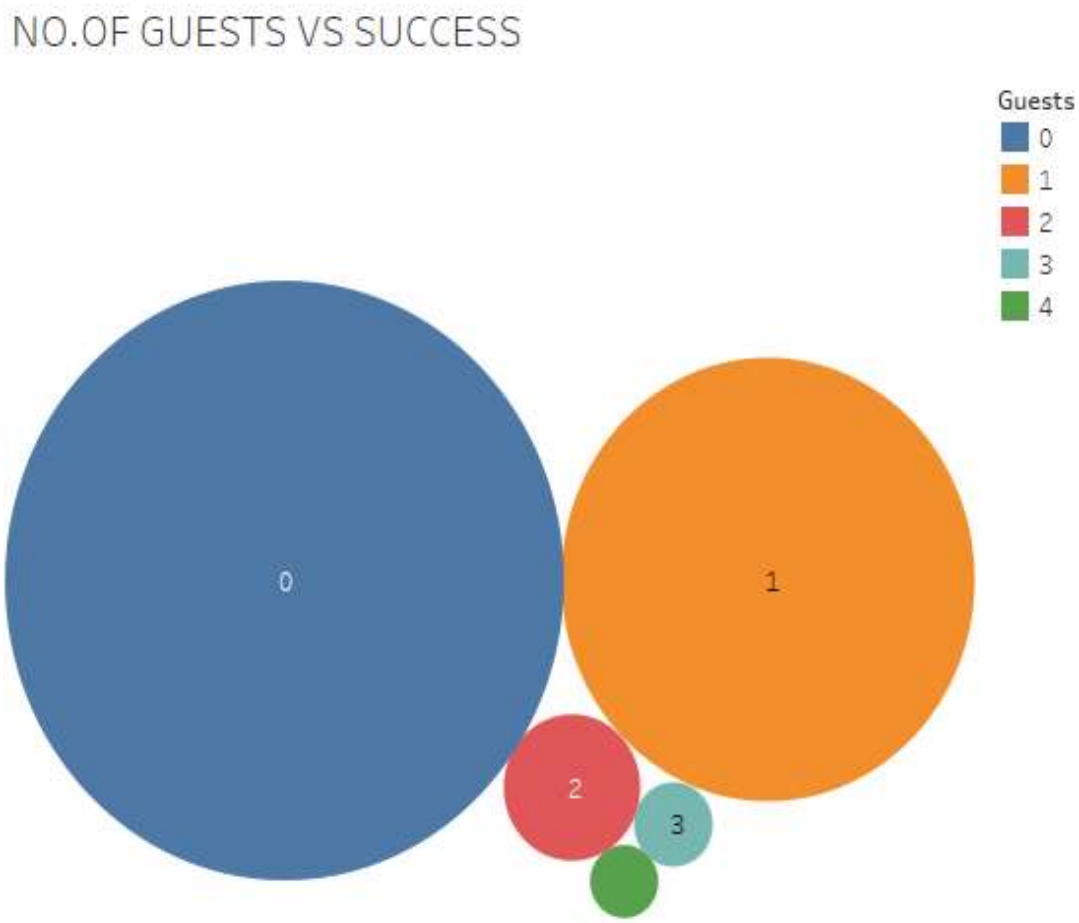
3. GENDER VS SUCCESS



Sum of Success for each Gender. Color shows details about count of Success. The data is filtered on Success, which keeps 1.

- FEMALES are having the most success rate.

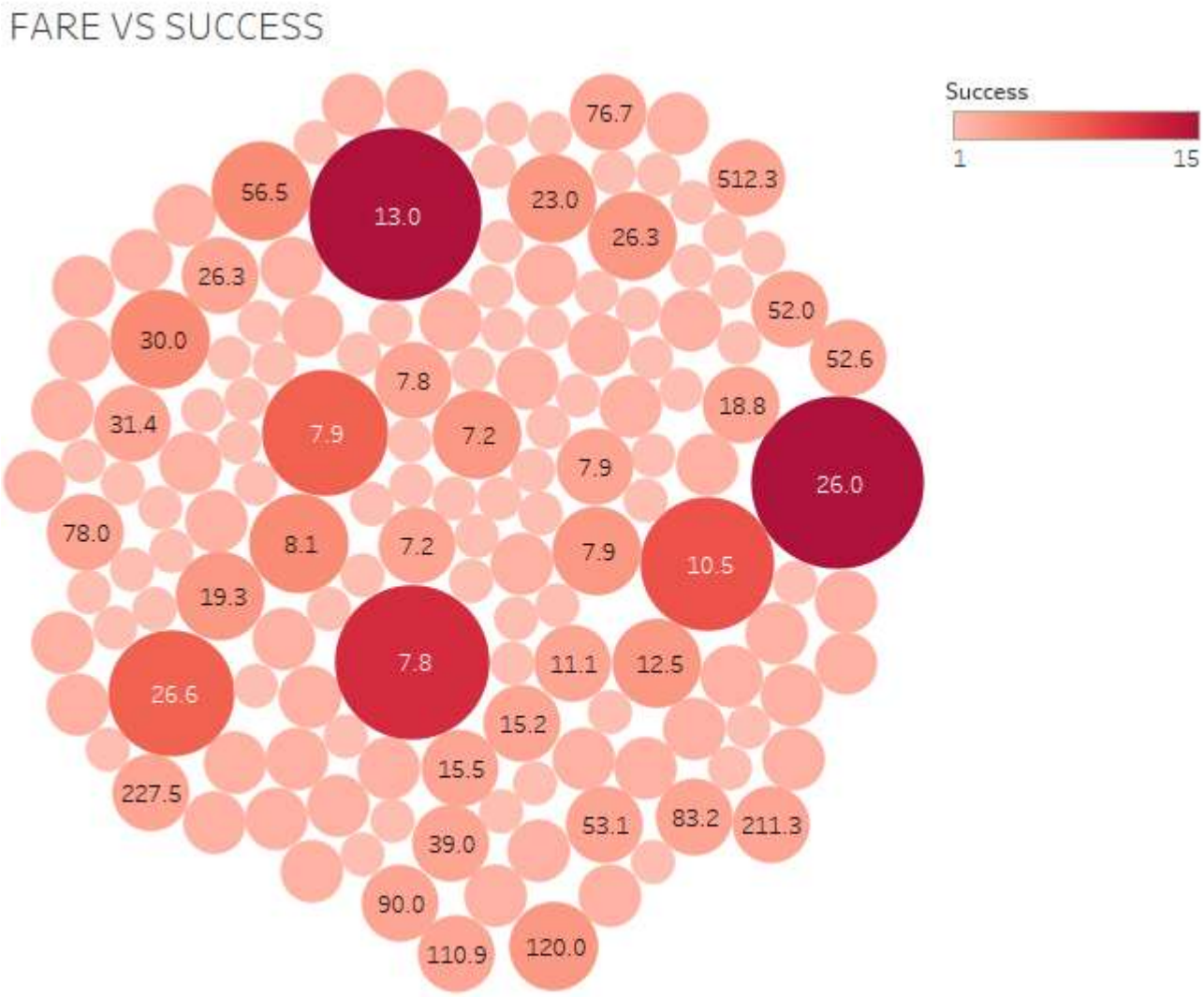
4.NO. OF GUESTS VS SUCCESS



Guests. Color shows details about Guests. Size shows count of Success. The marks are labeled by Guests. The data is filtered on Success, which keeps 1.

- Most People are travelling alone.
- With increase in no. of guest's success rate is decreasing

5. FARE VS SUCCESS



Fare. Color shows sum of Success. Size shows count of Success. The marks are labeled by Fare. The data is filtered on Success, which keeps 1.

- A Fare of around 26 or 13 is catchy to attract more customers
- With increase in the fare the

6 MILESTONE-4 (Attribute Preparation & Engineering for Mining Algorithm)

6.2 Loading CSV into weka

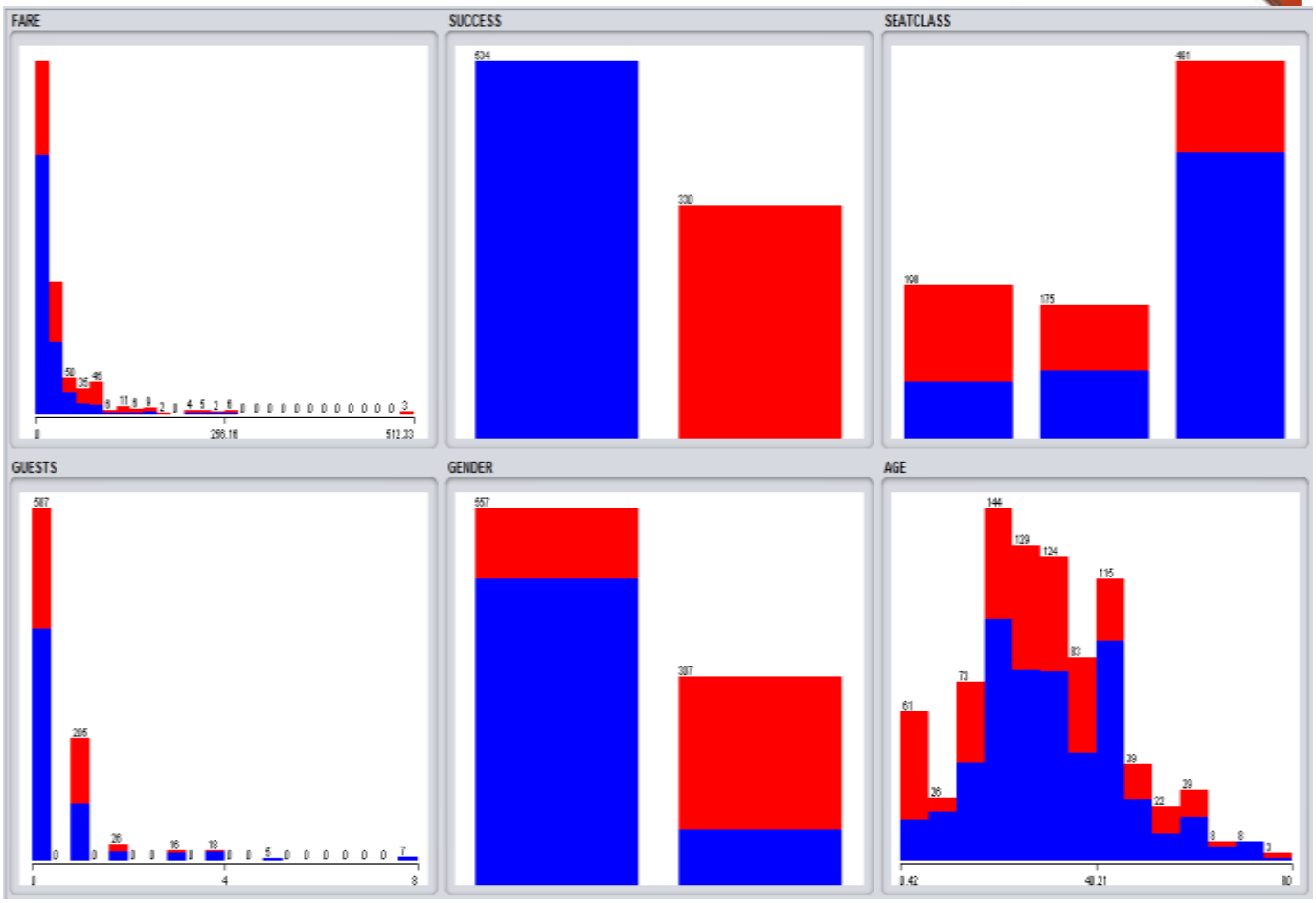
The obtained final csv file was loaded into weka. By default, all the attributes where changed to numeric type. But we have categorical attributes also so the type of SUCCESS and SEATCLASS was changed from numeric to nominal using the NumericToNominal filter.

6.3 Attribute selection

Attribute selection was done with Ranker Method &10-fold cross validation. The top 2 attributes were GENDER and FARE.

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 9 ===

average merit      average rank  attribute
 0.216 +- 0.005    1    +- 0      5 GENDER
 0.098 +- 0.006    2    +- 0      1 FARE
 0.085 +- 0.005    3    +- 0      3 SEATCLASS
 0.031 +- 0.004    4    +- 0      4 GUESTS
 0.017 +- 0.006    5    +- 0      6 AGE
```



The basic visualizations in weka for all the attributes with blue being failure and red being success.

7 MILESTONE-5 (Prediction Modeling and Visualization)

Modelling building was done in Weka with following three classification models being built. Although we identified top 2 significant attributes I have done modelling with all the attributes as we have only 5(excluding customer Id)

1. Decision Stump
2. Random Forest
3. J48

7.2 Decision Stump

A decision stump is a machine learning model consisting of a one-level decision tree. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature.

I used 10-fold cross validation to feed data.

Accuracy Measures

```
Correctly Classified Instances         679               78.588  %
Incorrectly Classified Instances       185               21.412  %
Kappa statistic                          0.5404
Mean absolute error                      0.3341
Root mean squared error                  0.4091
Relative absolute error                 70.7555 %
Root relative squared error             84.2032 %
Total Number of Instances              864
```

CONFUSION MATRIX

```
  a   b   <-- classified as
453  81 |   a = 0
104 226 |   b = 1
```

ACCURACY: 78.58%

Total no. of correctly classified instances is 679 out of 864

7.2 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

I used 10-fold cross validation to feed data.

Accuracy Measures

Correctly Classified Instances	707	81.8287 %
Incorrectly Classified Instances	157	18.1713 %
Kappa statistic	0.609	
Mean absolute error	0.2355	
Root mean squared error	0.3748	
Relative absolute error	49.8662 %	
Root relative squared error	77.1471 %	
Total Number of Instances	864	

CONFUSION MATRIX

```
  a   b   <-- classified as
469  65 |   a = 0
 92 238 |   b = 1
```

ACCURACY: 81.8287%

Total no. of correctly classified instances is 707 out of 864

7.2 J48

J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. It is most efficient classification algorithm with works like decision stump on each node it has splitted.

I used 10-fold cross validation to feed data.

Accuracy Measures

Correctly Classified Instances	708	81.9444 %
Incorrectly Classified Instances	156	18.0556 %
Kappa statistic	0.6113	
Mean absolute error	0.2469	
Root mean squared error	0.365	
Relative absolute error	52.2794 %	
Root relative squared error	75.1333 %	
Total Number of Instances	864	

CONFUSION MATRIX

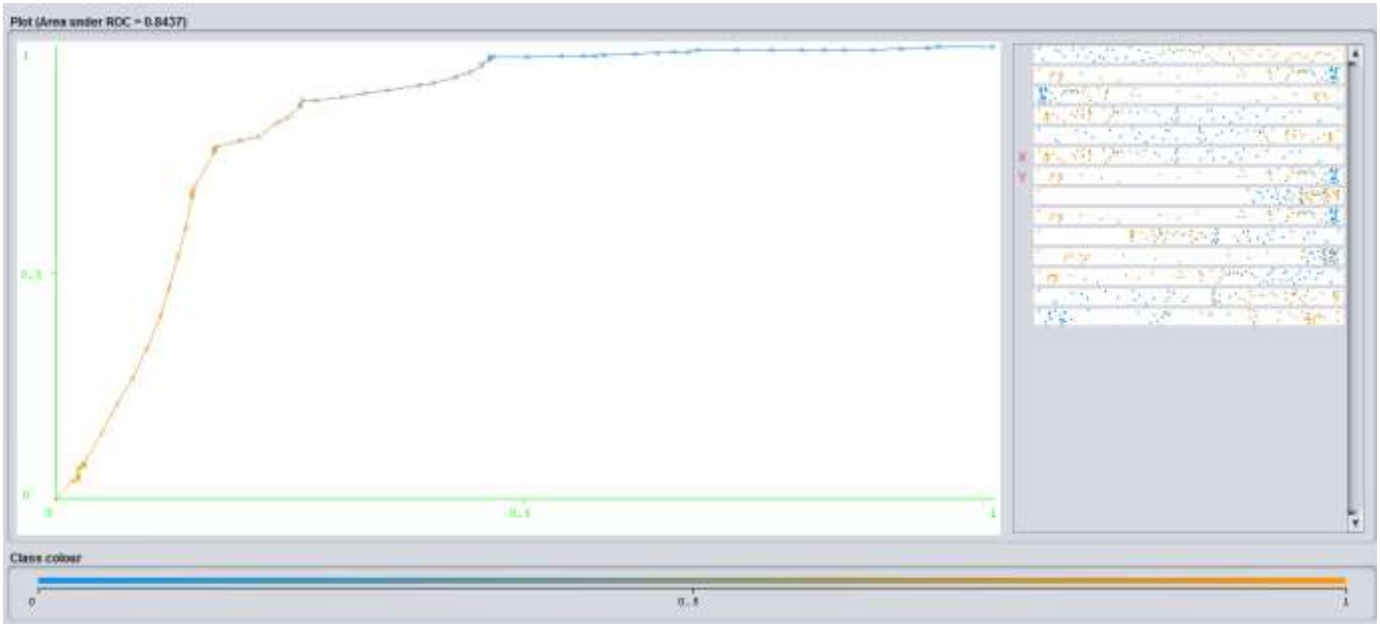
```
  a   b   <-- classified as
470  64 |   a = 0
 92 238 |   b = 1
```

ACCURACY: 81.944%

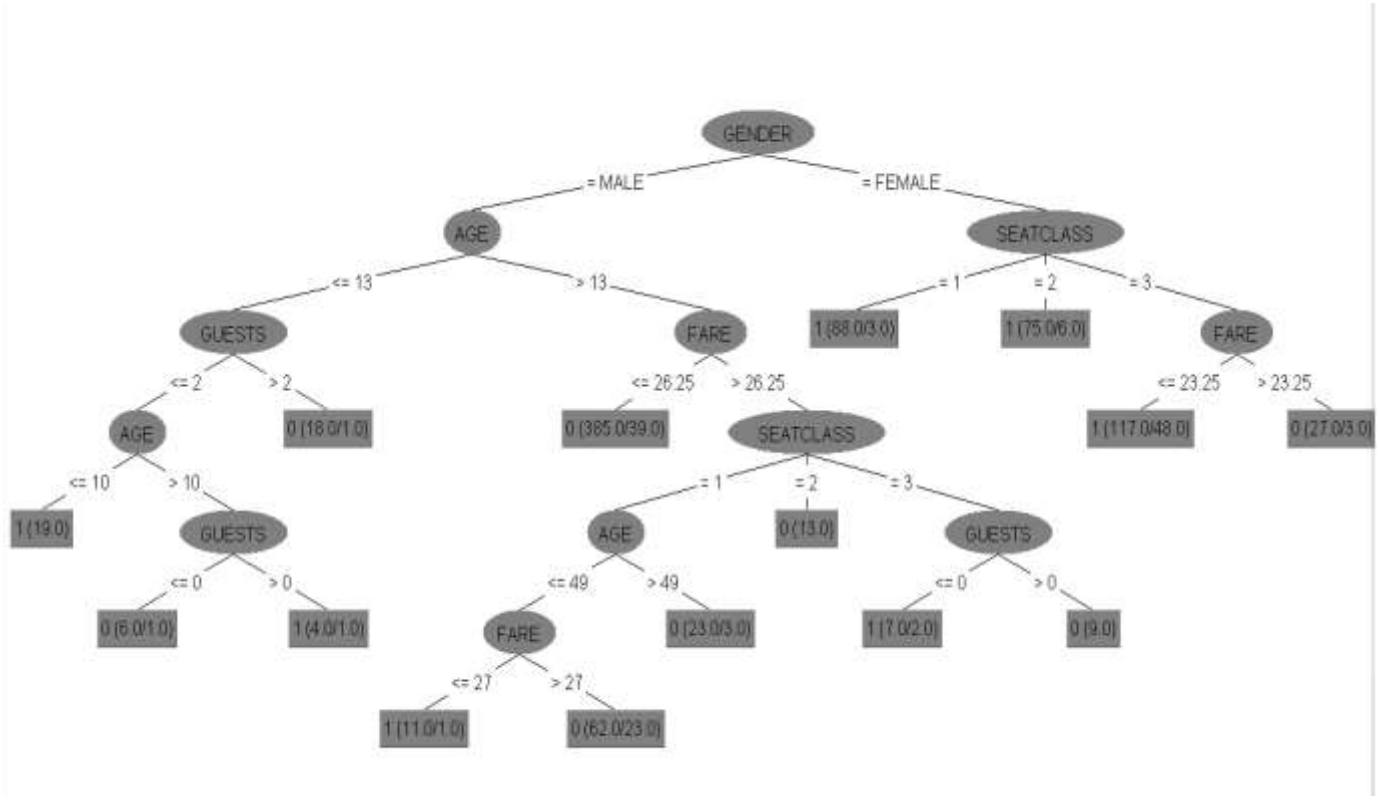
Total no. of correctly classified instances is 708 out of 864

Clearly based on accuracy of all three-decision stump, Random forest and J48 J48 is a better classifier.

ROC CURVE (0.8437)



TREE (PRIMARY NODE: GENDER)



8 INSIGHTS

- ✓ Provide offers in Seat Class 3 for customers with age greater than 23
- ✓ For customers with age greater than 13 have more offers for seat class 2
- ✓ Provide special package for bookings having customer between the age 10-13 and who are travelling with more than 2 guests

9 REFERENCES

1.Wikipedia
2.AIT 582 Lecture Slides by Professor Setareh Rafatirad

10 CODE

```
##Installing and Loading Required Packages
install.packages("RJSONIO")
install.packages("RCurl")
install.packages("sqldf")
install.packages("mice")
install.packages("VIM")
library(RJSONIO)
library(RCurl)
library(plyr)
library(stringr)
library(utils)
library(sqldf)
library(mice)
library(VIM)
library(lattice)
library(ggplot2)

##Getting Data in Json Format

raw_data <- getURL("http://ist.gmu.edu/~hpurohit/courses/ait582-proj-data-spring16.json")

##converting Data to CSV format

data <- fromJSON(raw_data)
length(data)
raw_data <- do.call(rbind, data)
raw_data=data.frame(raw_data,row.names = NULL)

##writing it to a csv file

write.csv(final_data, "raw_data.csv")

##Extracting Metadata from DESCRIPTION and assigning to AGE & GENDER

head(raw_data)
data=raw_data[-1,]
rownames(data) <- 1:nrow(data)
splitdat = do.call("rbind", strsplit(as.character(data$DESCRIPTION), ";"))
GENDER=str_extract(string =data$DESC,pattern = "(Mr | Miss | Mrs | Master)\\.")
mydata=cbind(data,GENDER,splitdat)
```

```
mydata[[9]] <- as.numeric(as.character(mydata[[9]]))
mydata=rename(mydata, c("1"="DESC","2"="AGE"))
mydata$DESCRIPTION=NULL
mydata$DESC=NULL
head(mydata)
sum(is.na(mydata$GENDER))

##Removing data whose GENDER is N/A

mydata=subset(mydata,!is.na(GENDER))
rownames(mydata) <- 1:nrow(mydata)

##Generating dataframe

Mr=sqldf("select * from mydata where GENDER='Mr.'")
Mrs=sqldf("select * from mydata where GENDER='Mrs.'")
Master=sqldf("select * from mydata where GENDER='Master.'")
Miss=sqldf("select * from mydata where GENDER='Miss.'")
mydata=sqldf("select * from Mr union select * from Mrs union select * from Master
union select * from Miss")

##Missing AGE imputation by Predictive Mean Matching(pmm) Seperatly for
each cateogry using SQL
#For Mr

Mr=sqldf("select * from mydata where GENDER='Mr.'")
temp_Mr=mice(Mr,m=1,method = 'pmm')
temp_Mr$imp$AGE
Mr=complete(temp_Mr,1)
Mr=sqldf(c("update Mr set GENDER='MALE'","select * from Mr"))

#For Mrs

Mrs=sqldf("select * from mydata where GENDER='Mrs.'")
temp_Mrs=mice(Mrs,m=1,method = 'pmm')
temp_Mrs$imp$AGE
Mrs=complete(temp_Mrs,1)
Mrs=sqldf(c("update Mrs set GENDER='FEMALE'","select * from Mrs"))

#For Master

Master=sqldf("select * from mydata where GENDER='Master.'")
temp_Master=mice(Master,m=1,method = 'pmm')
temp_Master$imp$AGE
```

```
Master=complete(temp_Master,1)
Master=sqldf(c("update Master set GENDER='MALE'","select * from Master"))

#For Miss

Miss=sqldf("select * from mydata where GENDER='Miss.'")
temp_Miss=mice(Miss,m=1,method = 'pmm')
temp_Miss$imp$AGE
Miss=complete(temp_Miss,1)
Miss=sqldf(c("update Miss set GENDER='FEMALE'","select * from Miss"))

#Combining all data

mydata=sqldf("select * from Mr union select * from Mrs union select * from Master
union select * from Miss")
write.csv(mydata, "mydata.csv")

#Writing data to load in weka&Tablue
mydata=read.csv("mydata.csv")
head(mydata)
mydata=mydata[,-1]
head(mydata)

#######BASIC STATISTICS**********

summary(mydata)
abc=sqldf("select * from mydata where FARE>200")
histogram(mydata$SUCCESS,mydata$SEATCLASS)
boxplot(mydata)
plot(mydata)


######################END##############################
###########
```