AIT-582

Application of Metadata in Complex Big Data Problems

Project Report

Fall-2017

K. Siva Naga Lakshmi

G01099587

Goal: Hands-on experience to extract, and utilize metadata in the data mining process
Scenario: You are a data scientist for an airline A, and you analyze a customer database. You want to identify the factors that are helpful to understand why some customers are flying your airline, and why others are canceling. Your data science team wants to recommend these factors to advertising team, such as demographic-specific packages to attract more customers.

Introduction

The Project has five milestones and they are described as follows:

1. Milestone - Data Acquisition and Conversion

In this step, the dataset is taken from the provided link. Using the R tool the dataset is converted from JSON to CSV File. Before converting the code, R packages need to be installed which are RCurl and RJSONIO.

Code (JSON to CSV) and deleting the extra header

582-lakshmi-airline.R* | airdata

Source on Save | Run | Source

```
1  library(RCurl)
2  library(RJSONIO)
3
4  airline.data=getURL("http://ist.gmu.edu/~hpurohit/courses/ait582-proj-data-spring16.json")
5  dataset=fromJSON(airline.data)
6  airline.data=do.call(rbind,dataset)
7  write.csv(airline.data,"airline.data.csv")
8
9  airdata=airline.data[-1,]
10 airdata=data.frame(airdata)
11 airdata$DESCRIPTION=as.character(airdata$DESCRIPTION)
```

Dataset

582-lakshmi-airline.R* | airdata

Filter

	FARE	DESCRIPTION	SUCCESS	SEATCLASS	GUESTS	CUSTOMERID
1	7.25	Braund, Mr. Owen Harris;22	0	3	1	1
2	71.2833	Cumings, Mrs. John Bradley (Florence Briggs Thayer);38	1	1	1	2
3	7.925	Heikkinen, Miss. Laina;26	1	3	0	3
4	53.1	Futrelle, Mrs. Jacques Heath (Lily May Peel);35	1	1	1	4
5	8.05	Allen, Mr. William Henry;35	0	3	0	5
6	8.4583	Moran, Mr. James;	0	3	0	6
7	51.8625	McCarthy, Mr. Timothy J;54	0	1	0	7
8	21.075	Palsson, Master. Gosta Leonard;2	0	3	3	8
9	11.1333	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg);27	1	3	0	9
10	30.0708	Nasser, Mrs. Nicholas (Adele Achem);14	1	2	1	10
11	16.7	Sandstrom, Miss. Marguerite Rut;4	1	3	1	11

2. Milestone - Metadata Extraction and Imputation

The main objective of this milestone is to identify the metadata types. With “Description” data field, the data can be extracted and append them as additional data columns. The extracted metadata fields are as following Gender, age, first name, last name and the prefix.

The above metadata fields have been appended as separate columns using the R functions ‘stringr’, ‘grepl’ and ‘sapply’. In the Gender column data has been derived using the prefixes “Mr. /Master” considered as male and female for all the remaining values.

SUCCESS	SEATCLASS	GUESTS	CUSTOMERID	LastName	prefix	FirstName	Gender	Age
0	3	1	1	Braund	Mr	Owen Harris	Male	22.00
1	1	1	2	Cumings	Mrs	John Bradley (Florence Briggs Thayer)	Female	38.00
1	3	0	3	Heikkinen	Miss	Laina	Female	26.00
1	1	1	4	Futrelle	Mrs	Jacques Heath (Lily May Peel)	Female	35.00
0	3	0	5	Allen	Mr	William Henry	Male	35.00
0	3	0	6	Moran	Mr	James	Male	NA
0	1	0	7	McCarthy	Mr	Timothy J	Male	54.00
0	3	3	8	Palsson	Master	Gosta Leonard	Male	2.00
1	3	0	9	Johnson	Mrs	Oscar W (Elisabeth Vilhelmina Berg)	Female	27.00
1	2	1	10	Nasser	Mrs	Nicholas (Adele Achem)	Female	14.00
1	3	1	11	Sandstrom	Miss	Marguerite Rut	Female	4.00
1	1	0	12	Bonnell	Miss	Elizabeth	Female	58.00
0	3	0	13	Saundercock	Mr	William Henry	Male	20.00
0	3	1	14	Andersson	Mr	Anders Johan	Male	39.00
0	3	0	15	Vestrom	Miss	Hulda Amanda Adolfina	Female	14.00
1	2	0	16	Hewlett	Mrs	(Mary D Kingcome)	Female	55.00
0	3	4	17	Rice	Master	Eugene	Male	2.00

There were missing values in the age column, Using the mean imputation method NA values has been replaced with mean value of “29.56”, using the floor function in R it has been rounded off to the nearest value.

After imputing the missing values in the age column, one more column is added here where the age category is defined. Now the data is ready to use for further analysis.

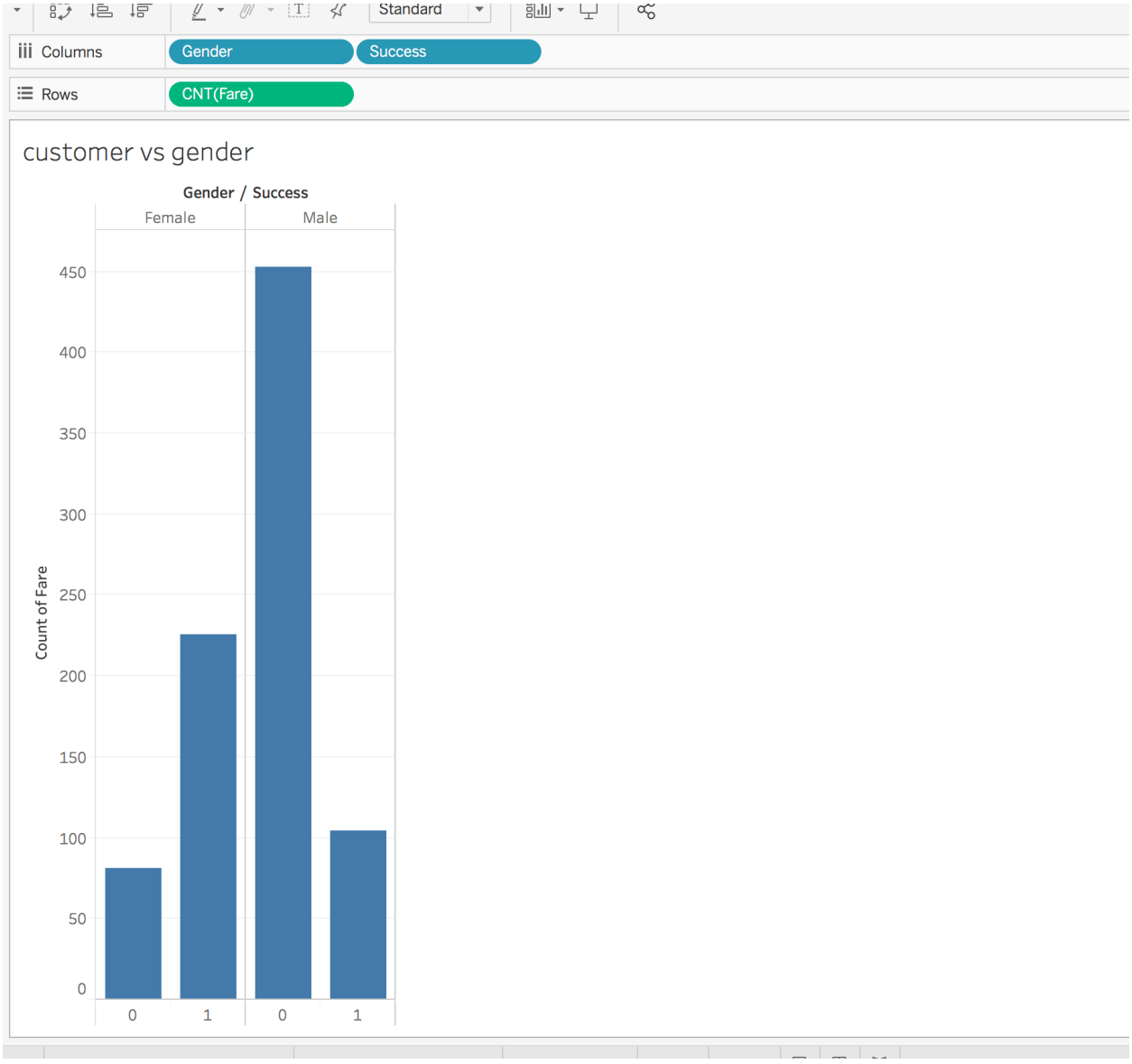
A	B	C	D	E	F	G	H	I	J	K	L
	FARE	SUCCESS	SEATCLASS	GUESTS	CUSTOMERID	LastName	prefix	FirstName	Gender	Age	agecategory
1	7.25	0	3	1	1	Braund	Mr	Owen Harris	Male	22	Young
2	71.2833	1	1	1	2	Cumings	Mrs	John Bradley (Florence Briggs Tha	Female	38	Middle Aged
3	7.925	1	3	0	3	Heikkinen	Miss	Laina	Female	26	Young
4	53.1	1	1	1	4	Futrelle	Mrs	Jacques Heath (Lily May Peel)	Female	35	Middle Aged
5	8.05	0	3	0	5	Allen	Mr	William Henry	Male	35	Middle Aged
6	8.4583	0	3	0	6	Moran	Mr	James	Male	29	Young
7	51.8625	0	1	0	7	McCarthy	Mr	Timothy J	Male	54	Middle Aged
8	21.075	0	3	3	8	Palsson	Master	Gosta Leonard	Male	2	Infant
9	11.1333	1	3	0	9	Johnson	Mrs	Oscar W (Elisabeth Vilhelmina Ber	Female	27	Young
10	30.0708	1	2	1	10	Nasser	Mrs	Nicholas (Adele Achem)	Female	14	Teenage
11	16.7	1	3	1	11	Sandstrom	Miss	Marguerite Rut	Female	4	minor
12	26.55	1	1	0	12	Bonnell	Miss	Elizabeth	Female	58	Middle Aged
13	8.05	0	3	0	13	Saundercock	Mr	William Henry	Male	20	Young
14	31.275	0	3	1	14	Andersson	Mr	Anders Johan	Male	39	Middle Aged
15	7.8542	0	3	0	15	Vestrom	Miss	Hulda Amanda Adolfina	Female	14	Teenage
16	16	1	2	0	16	Hewlett	Mrs	(Mary D Kingcome)	Female	55	Middle Aged
17	29.125	0	3	4	17	Rice	Master	Eugene	Male	2	Infant
18	13	1	2	0	18	Williams	Mr	Charles Eugene	Male	29	Young
19	18	0	3	1	19	Vander Plank	Mrs	Julius (Emelia Maria Vandemoorte	Female	31	Middle Aged
20	7.225	1	3	0	20	Masselmani	Mrs	Fatima	Female	29	Young
21	26	0	2	0	21	Fynney	Mr	Joseph J	Male	35	Middle Aged

Above screenshot has shown that data is ready to be analyzed for further steps.

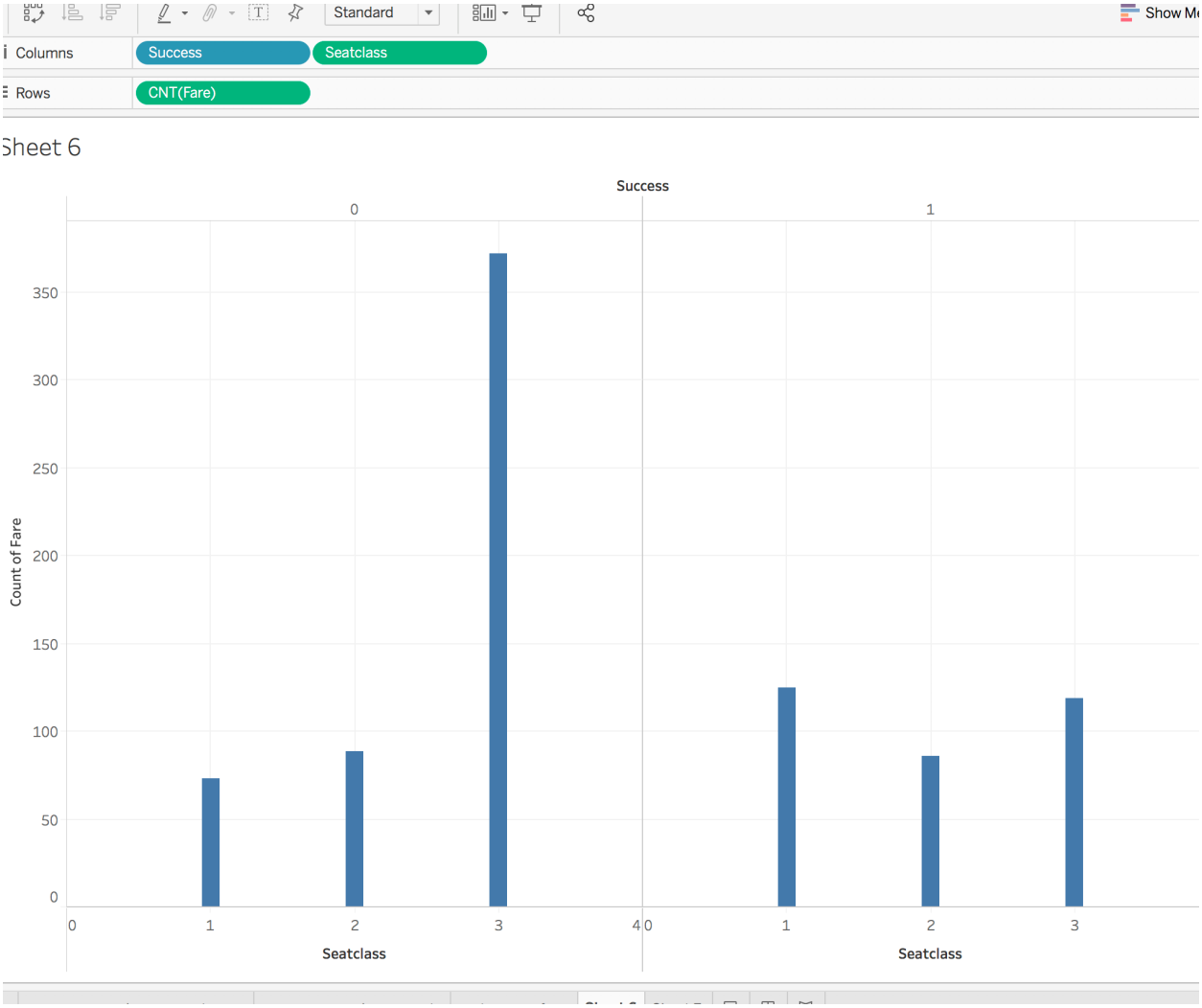
3. Metadata Exploration

In this milestone visualization of pattern between different metadata fields using the “Tableau” software and the results are interpreted.

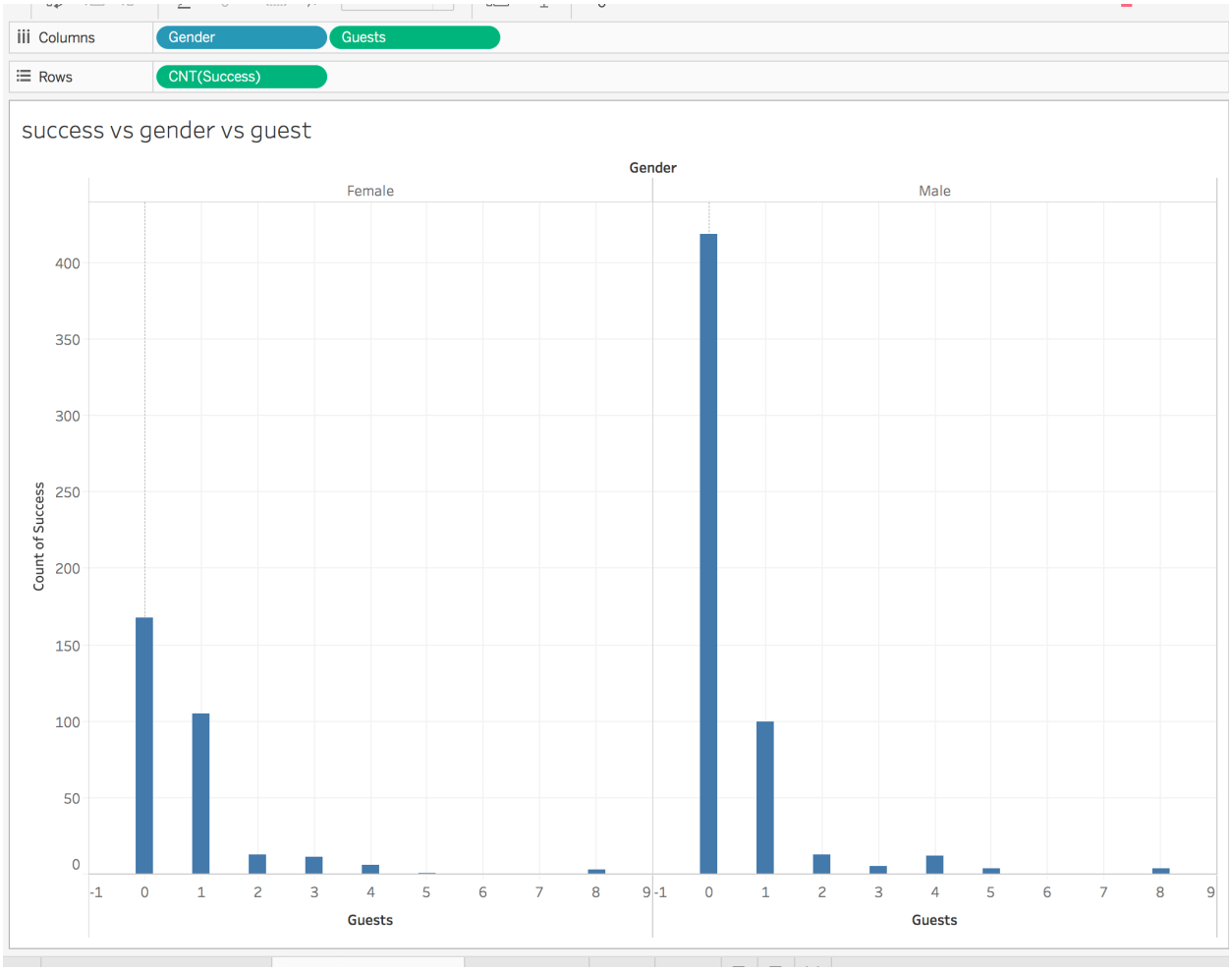
a) Gender vs Fare with Success- From the graph we can observe that males have the highest number of cancellations.



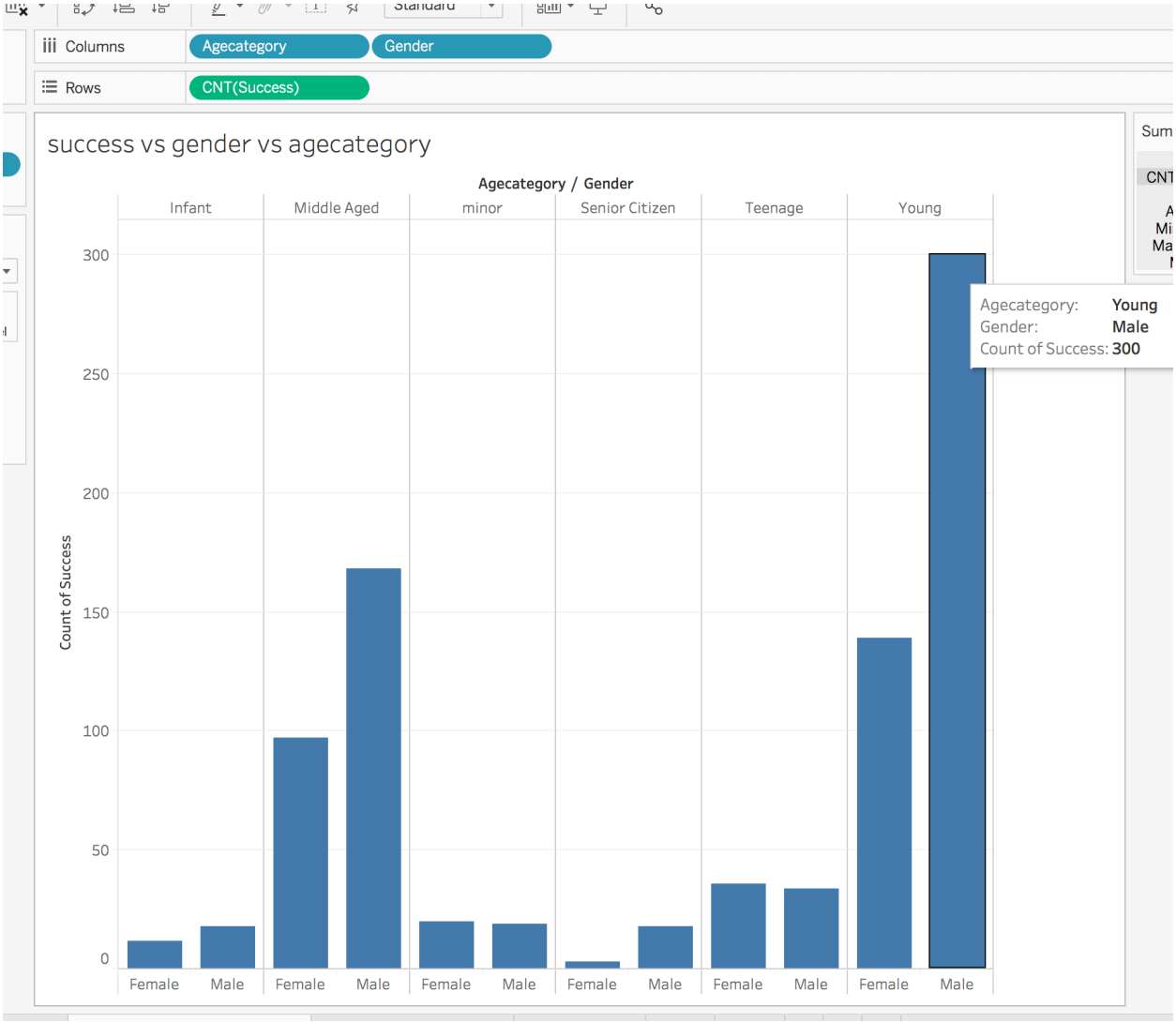
b) Success vs Fare with seat class- from the graph we can observe that there are lot of cancellations in the 3rd seat class.



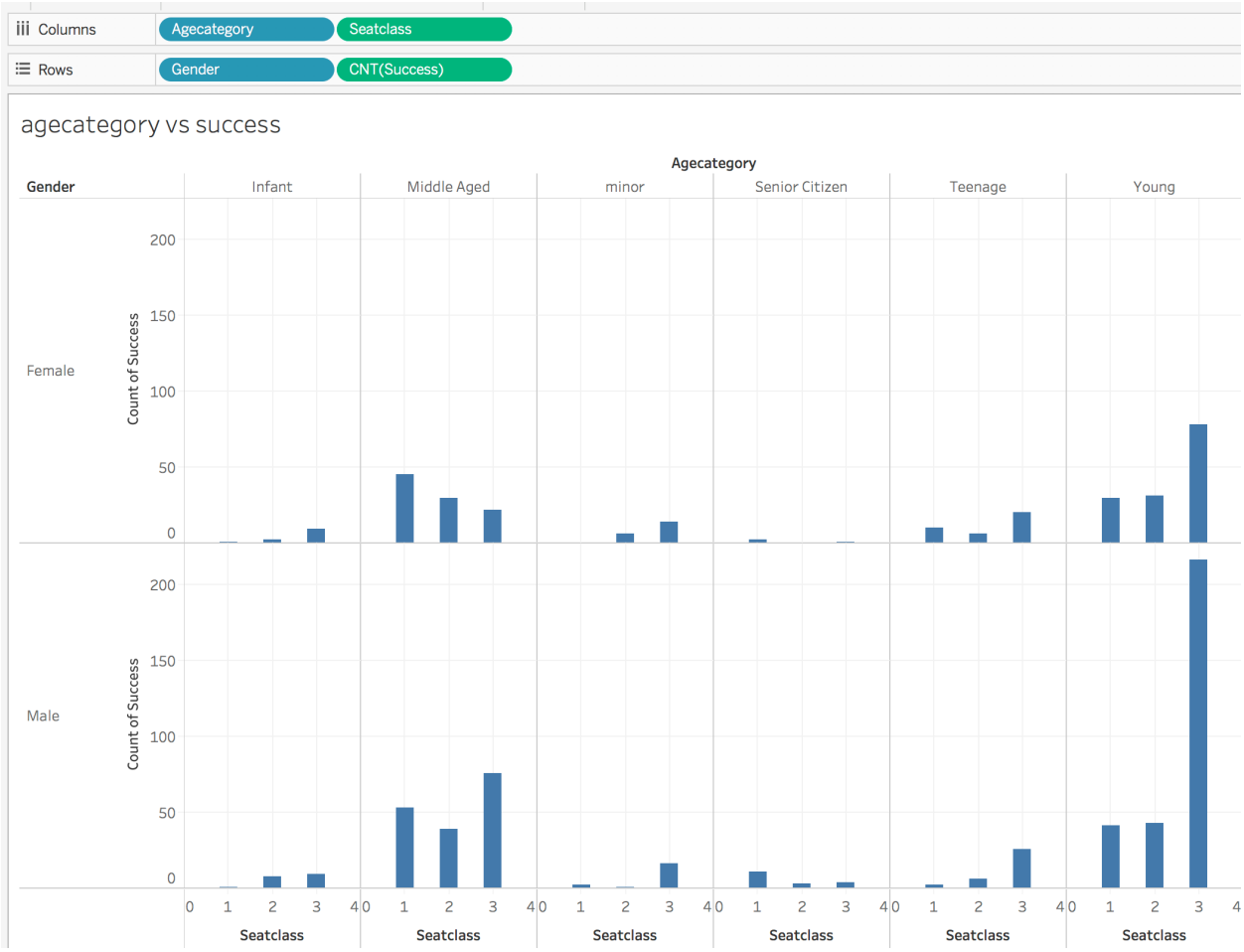
c) Guest vs Success with gender- From the graph we can observe that, males are travelling more with 0 guest when compare to female.



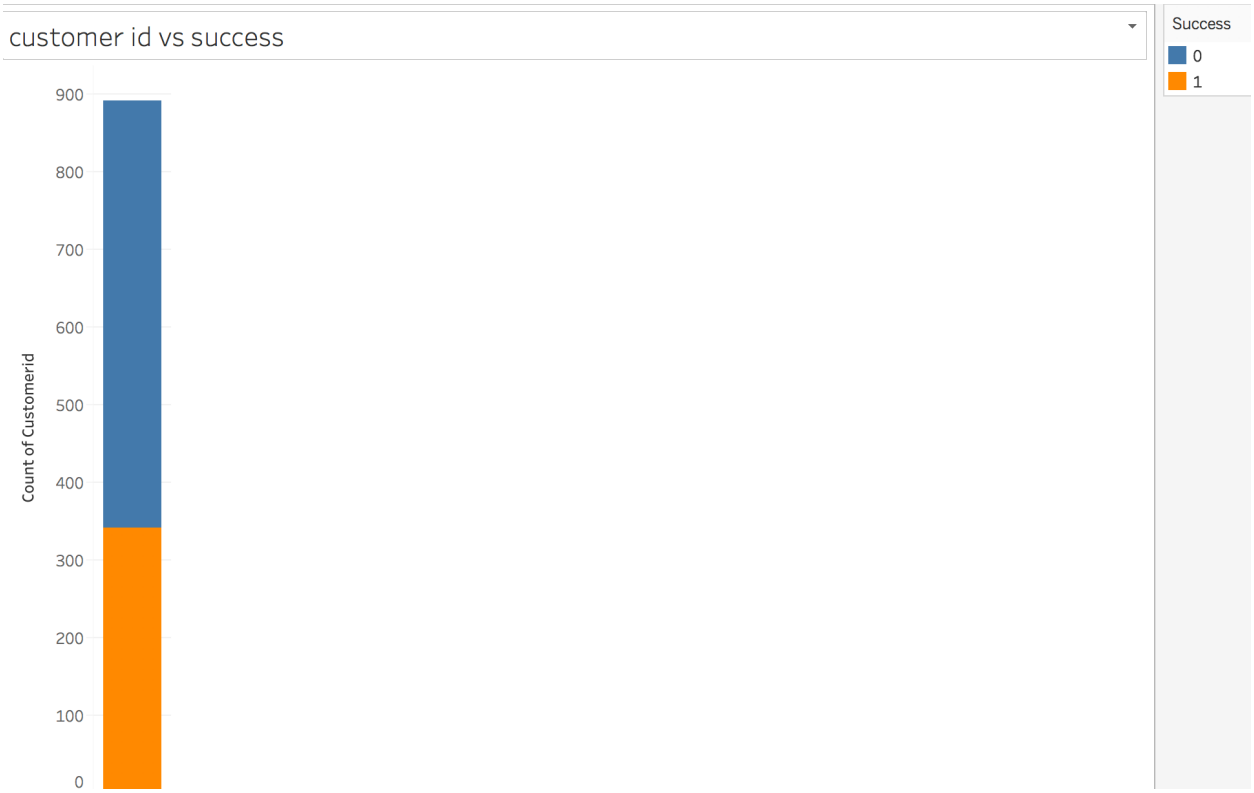
d) Success vs Gender with Age category- From the graph we can observe that, the count of success is high in male and in the young category compared to female. Second category would be the middle aged.



e) Success vs Seat class with Age Category and Gender- From the graph we can observe that male count is high in the seat-class 3where the young people preferred more.

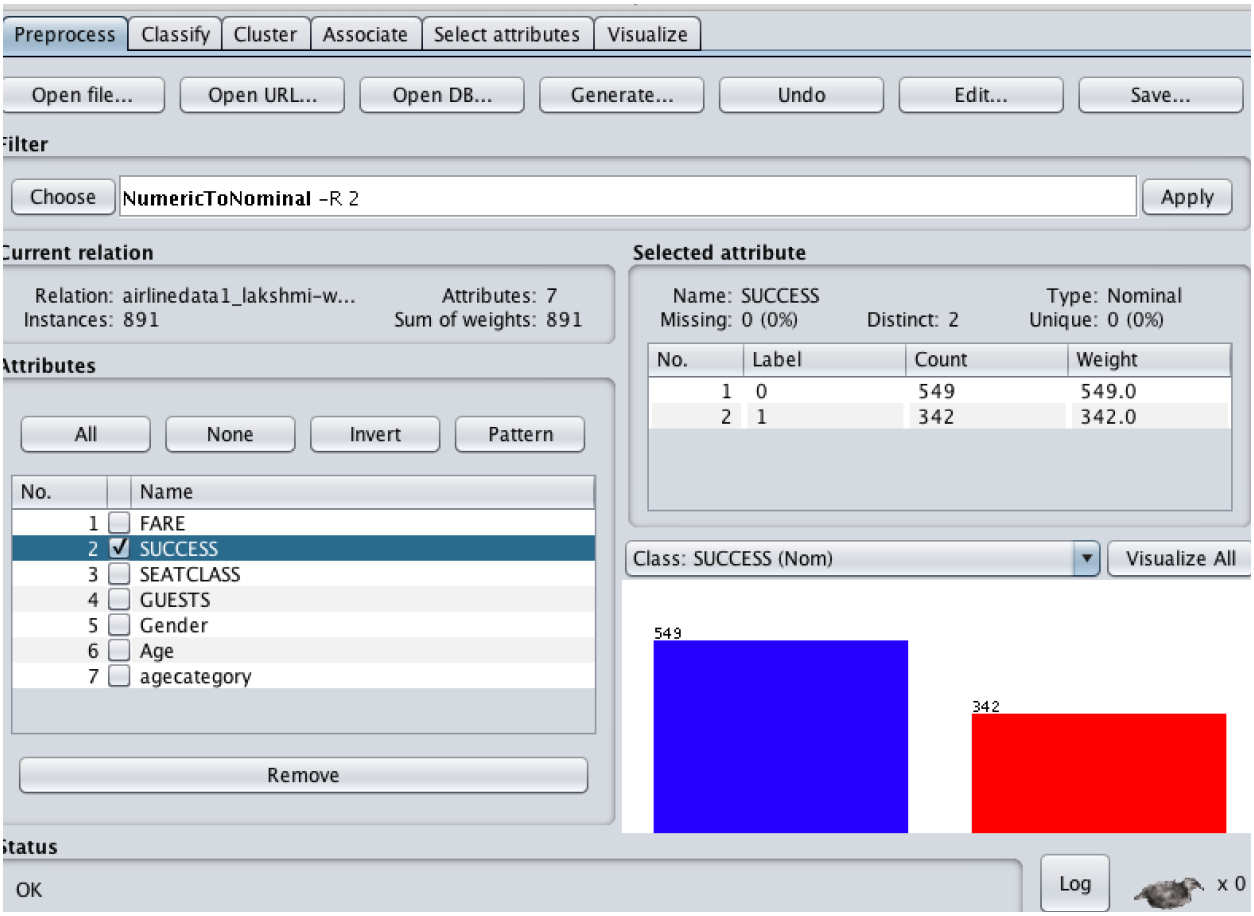


f) Customer ID vs Success- from the graph we can observe that where 0 indicates persons not willing to fly, and 1 indicates people willing to fly. Count of customers with 0 is 549 and 1 is 342.



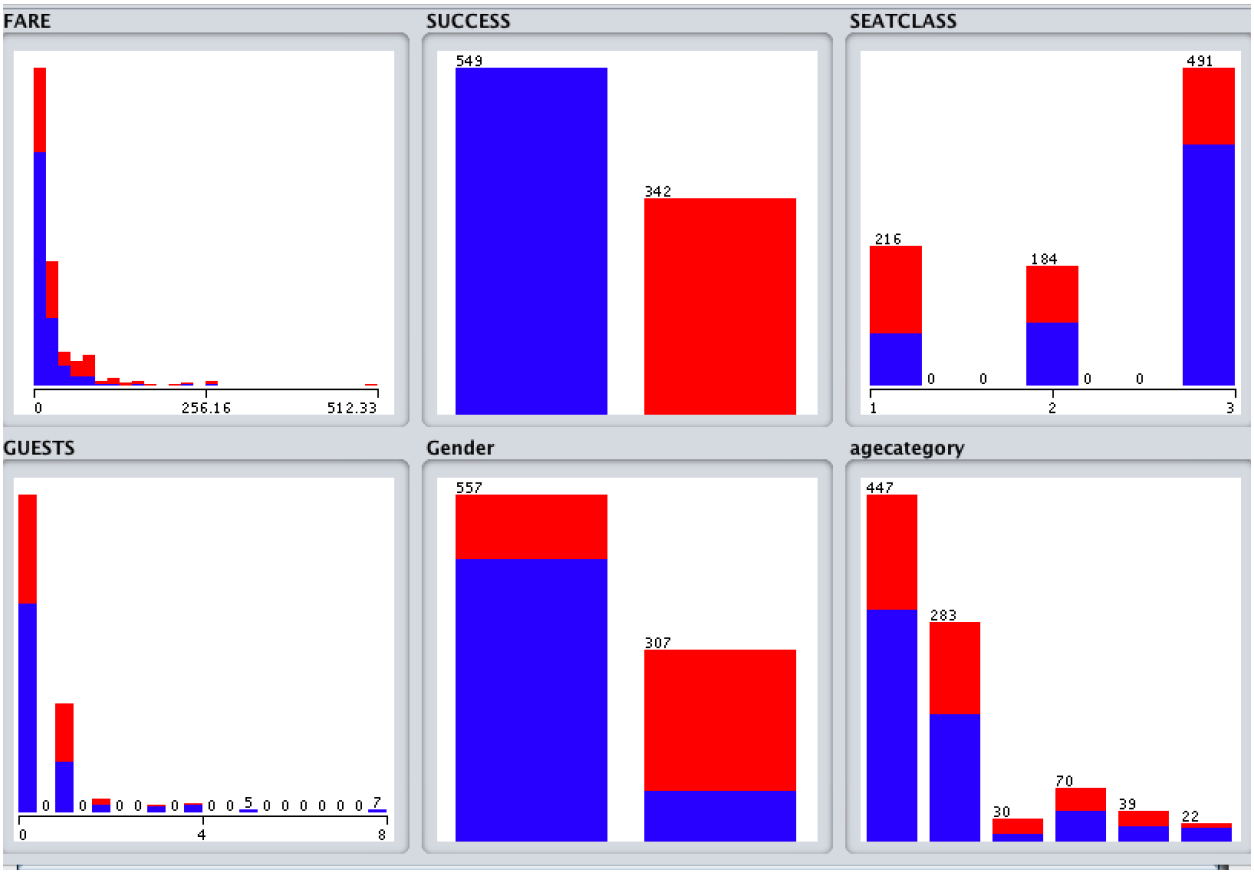
4. Attribute Preparation and Engineering

In this milestone, the extracted metadata file is converted to ARFF using the Weka tool. Loading the csv file in the ARFF viewer and saved it by using the ARFF extension.

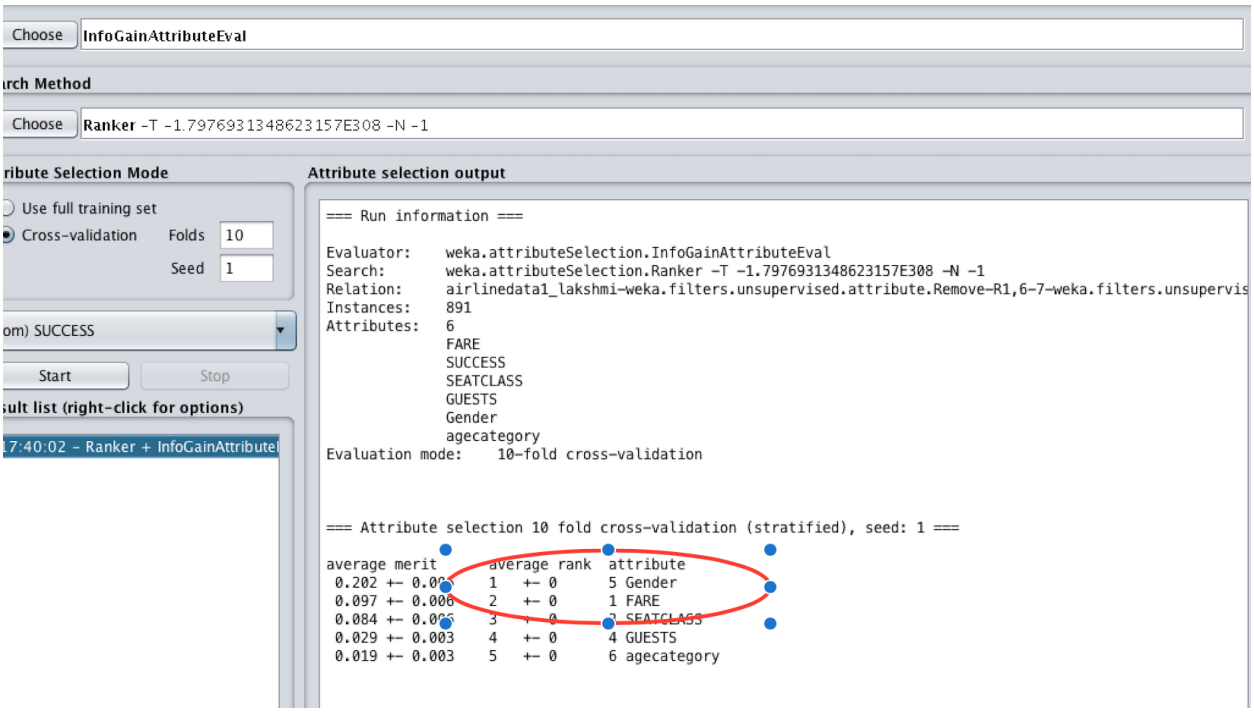


The above screenshot is the preprocessing step, in this data file is imported to Weka tool by clicking open file and giving the corresponding path where data file is located. After data is loaded, attribute type is changed to numeric to nominal using filter option. The result matches with the tableau one's in the Weka tool.

Similarly plotting other graphs can be done by clicking on the ‘visualize all’, where it displays all the graphs.



Attribute selection, here the objective of this step is to find the top attributes using 10-fold cross validation model. Where I opted the ranker method to find out the top attributes, which in-turn automatically takes the info gain attribute evaluation. Using this method, the top two attributes will be the Gender and the Fare. Using these two attributes, all other values or fields can be classified and they act as roots. Results are shown below in the screen shots

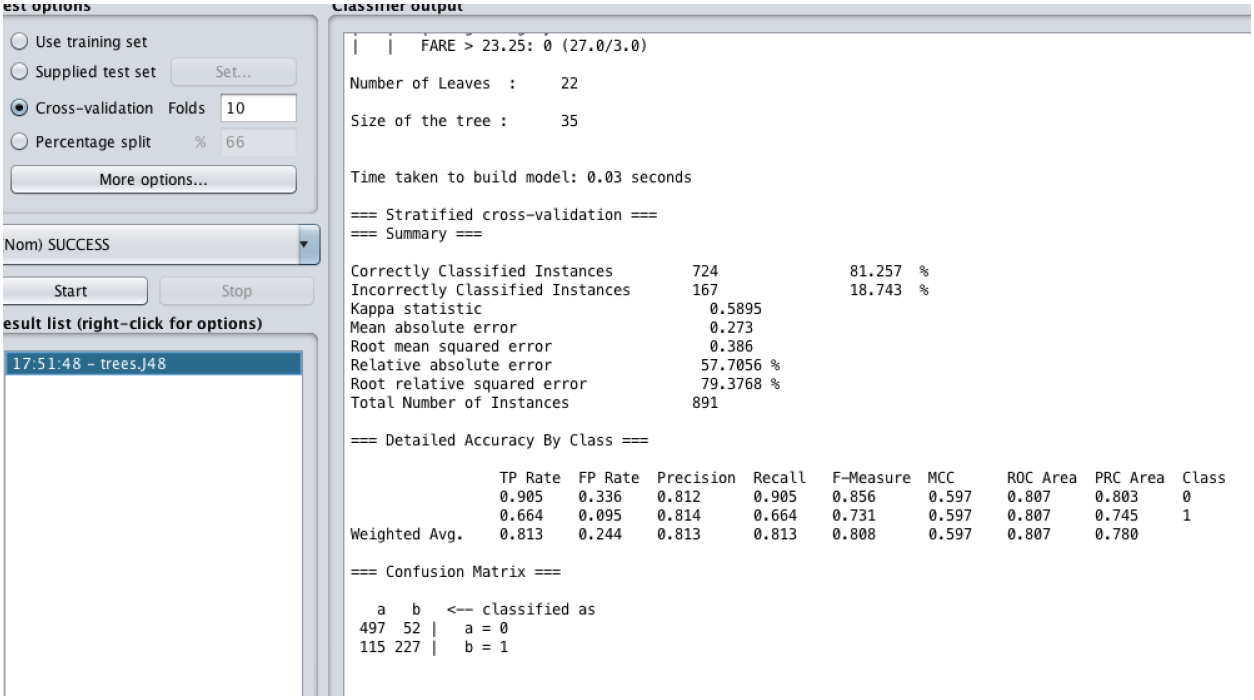


In the above screen shot, where gender, fare is represented in the eclipse and ranked as top two attributes.

5. Prediction Modeling and Visualization

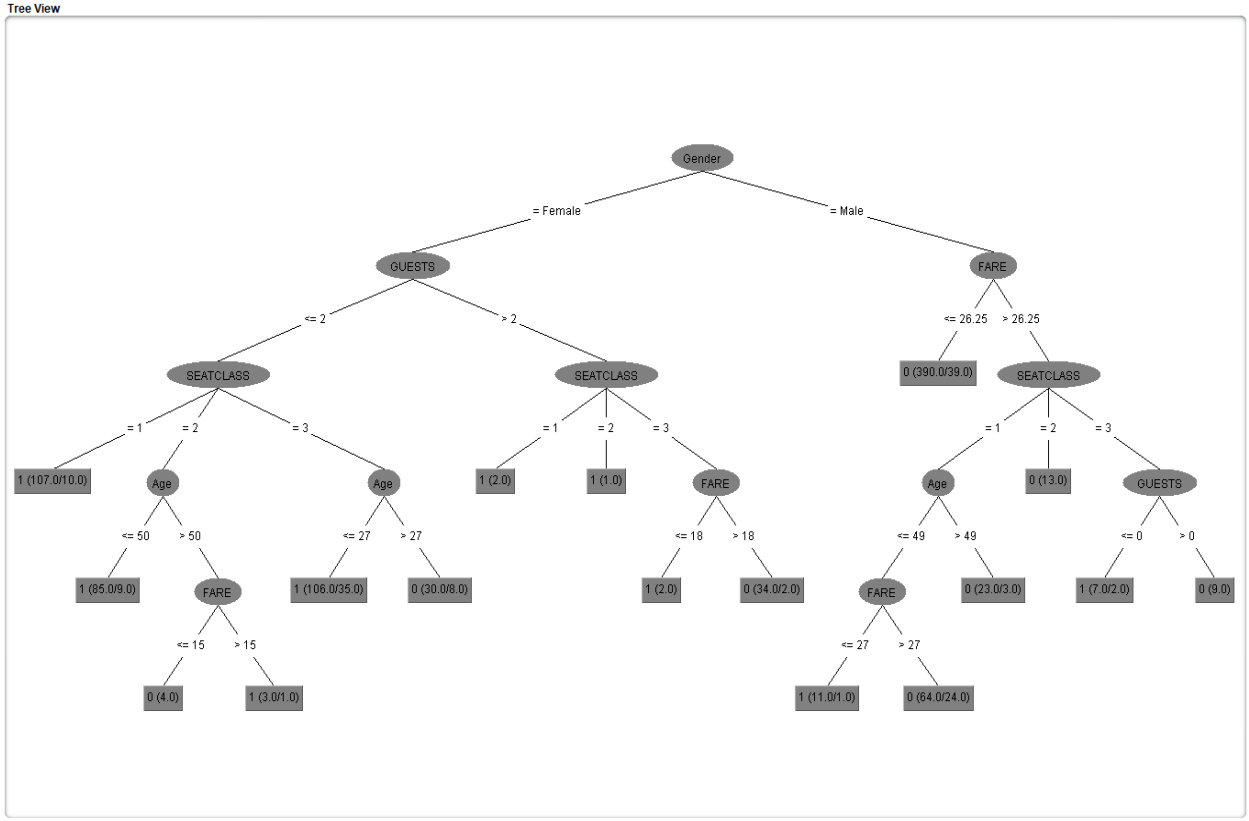
Objective of this milestone is to design a classification model using the J48 (decision tree) and Random forest algorithms. And also, to generate the ROC curve using 10-fold cross validation method.

For classification model, the data is divided into the training and the testing data by using the default classification feature of Weka. Data is divided into training as 80% and 20% test data. After running the **J48 model** on the training data the results are obtained as below in the screen shot. From the picture, we can say that accuracy of this model is **81.257%** with **724** correctly classified instances.



Above screen shot depicts classification has done using **J48 algorithm**

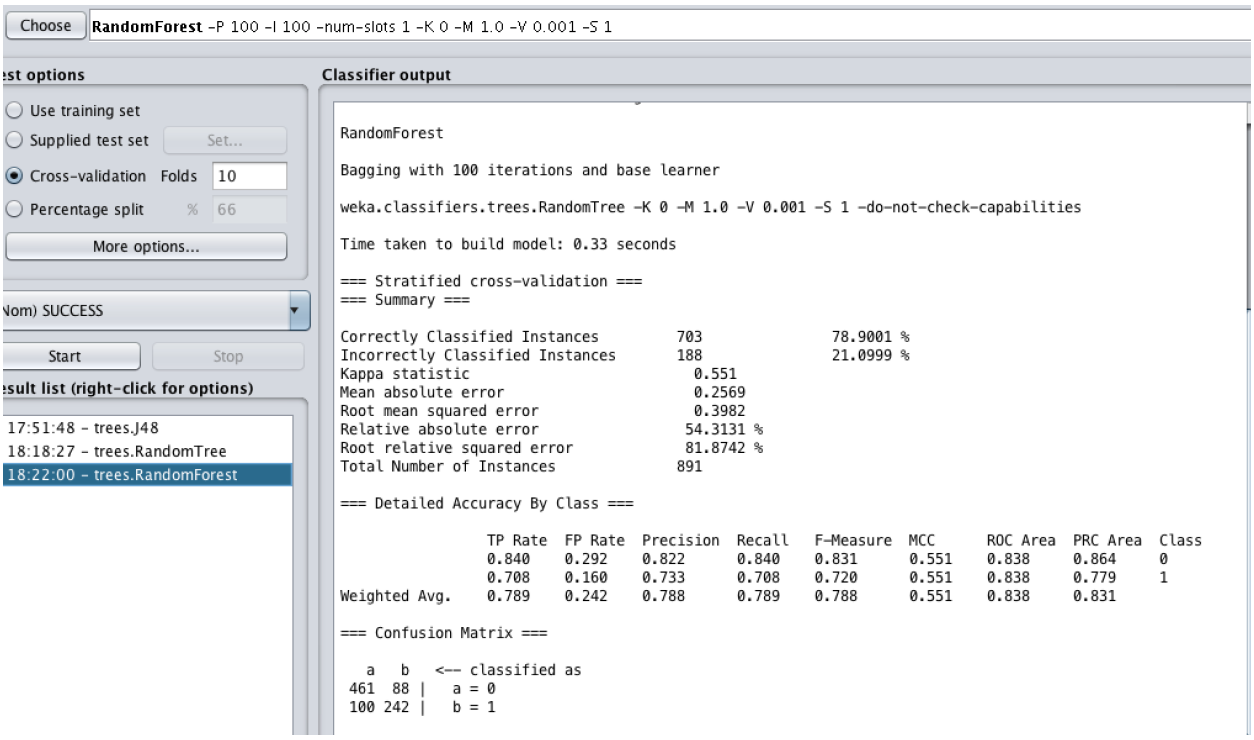
Tree View



Decision Tree

After the Decision tree, I used the **Random Forest** method to classify the model and to check the significance of the method. The accuracy of this model gives as **78.90%** with **703** as correctly classified instances.

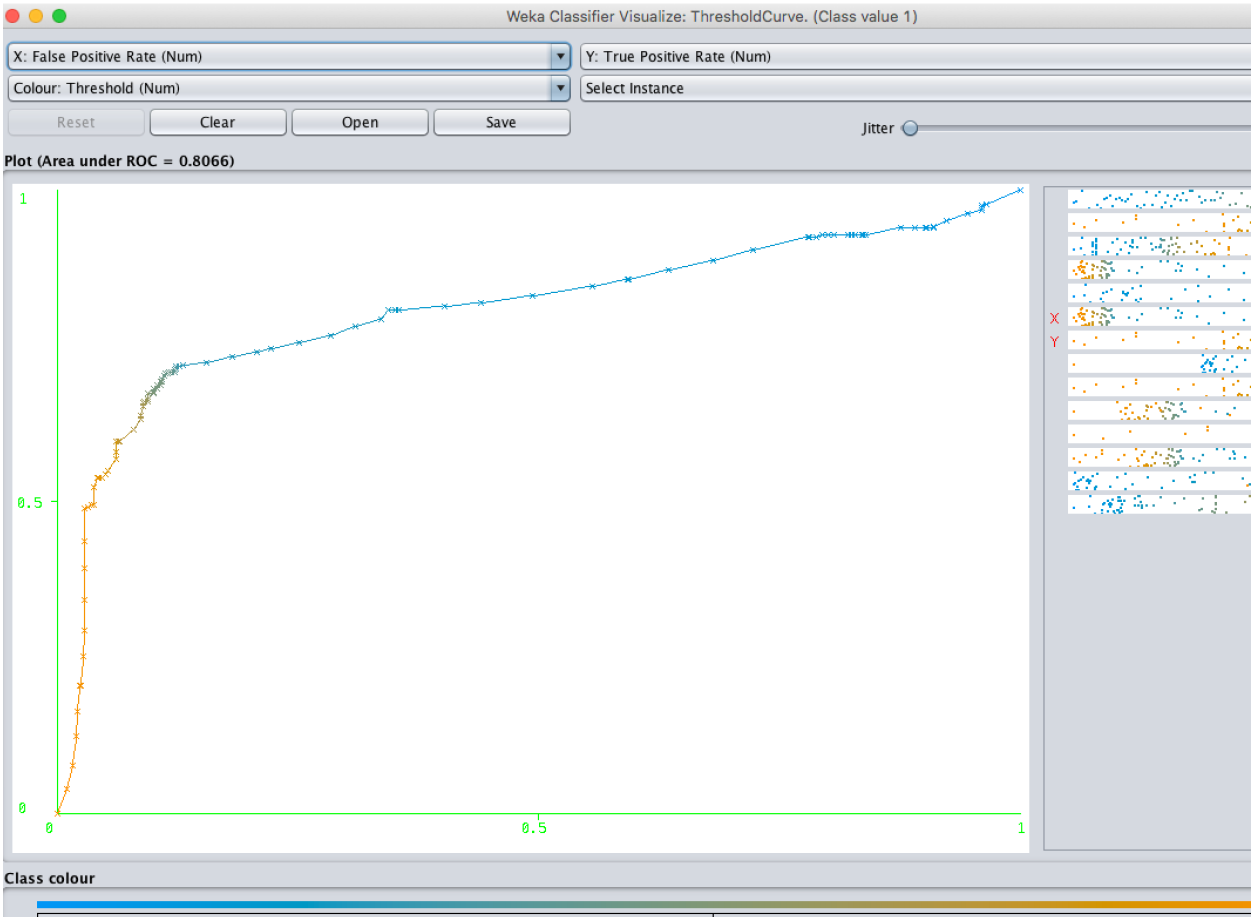
Random Forest Screenshot



ROC Curves

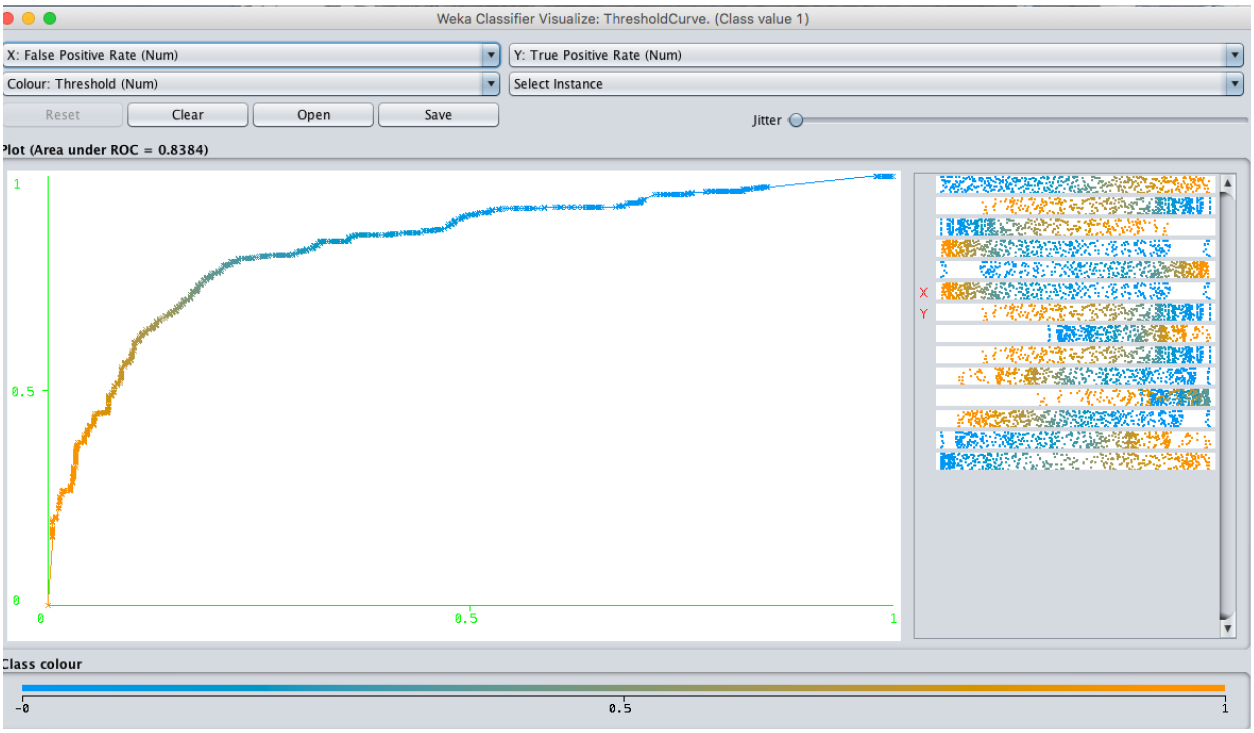
It gives the graphical illustration of the performance of a binary classifier system as its discrimination threshold is varied. Plotting the ROC curve is done by decision tree and random forest algorithms using 10-fold cross validation.

a) ROC-Decision Tree(J48)
For the threshold value 1 as shown below. The area under the curve is **0.8066**, which means it is a good model and if the curve is leaning towards more to the y-axis then it is considered as good model and more accurate.



b) ROC-Random Forest

For threshold value 1 is shown below, from the graph we can observe that area under curve is **0.834.**



Area under curve is more compared to decision tree. Accuracy is more for decision tree model and also the curve for random forest is smooth than decision tree model, but decision model is more better than random forest.

If we increase the true positive values, the curve will be more accurate and it will be towards the y-axis. When compared both the models, decision tree has truer positive value. The model will be robust.

Insights

Few insights can be drawn from the milestones

- Top Two attributes are the Gender and Fare where we can predict the values using these two factors.
- Males has higher success rate compared to females.
- The young age and middle age customers are more likely to travel.

If company works focuses on the above factors, it will be able to get more sales and the success rate will increase.