Airline Project Report

Shravan Chintha
G01064991

For the individual project, I have chosen to work on Airline data set. The goal of this project is to extract metadata and use the metadata in creating a classification or prediction model for a problem that given a customer will fly or not fly and based on the prediction model, recommend these factors to advertising team to design customer demographic specific packages to attract more customers.

Project Milestones:

There are five milestones to complete this scenario.

i. Data Acquisition and Conversion
ii. Metadata Extraction and Imputation
iii. Metadata Exploration
iv. Attribute Preparation and Engineering for preparing for Mining Algorithm
v. Prediction Modelling and Visualization

The above milestones have been implemented in step by step process to achieve the goal. Results of each milestone are documented in this report.

i. Data Acquisition and Conversion

The objective of this step is to download the given project data file from http://ist.gmu.edu/~hpurohit/courses/ait582-proj-data-spring16.json programmatically and convert the downloaded JSON file to CSV format for easier manipulations.

Downloaded the data file using R script. The downloaded data file is in JSON format, hence it is to be converted into CSV file to make manipulations easy.

JSON file is converted to CSV using R script and deleted extra row which repeats the header to make the data look good for further manipulations and interpretations. The CSV data file thus obtained looks as shown below:

	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
12	26.55	Bonnell, Miss. Elizabeth;58	1	1	0	12
13	8.05	Saundercock, Mr. William Henry;20	0	3	0	13
14	31.275	Andersson, Mr. Anders Johan;39	0	3	1	14
15	7.8542	Vestrom, Miss. Hulda Amanda Adolfina;14	0	3	0	15
16	16	Hewlett, Mrs. (Mary D Kingcome) ;55	1	2	0	16
17	29.125	Rice, Master. Eugene;2	0	3	4	17
18	13	Williams, Mr. Charles Eugene;	1	2	0	18
19	18	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortel...	0	3	1	19
20	7.225	Masselmani, Mrs. Fatima;	1	3	0	20

ii. Metadata Extraction and Imputation

The main objective of this milestone is to identify metadata types in the data field “Description”, extract those metadata fields and append them as additional fields to each of the data record. Also, as there would be missing fields or data to some of the records, those missing values should be imputed using any of the imputing methods.

Identified the below fields as additional metadata in the column “Description”:

- First Name
- Last Name
- Age
- Gender (from title “Mr.”, “Mrs.” Etc.)

Extracted the above-mentioned metadata fields and appended the fields as separate columns to each data record using “stringsplit”, “grepl”, “sapply” functions in R. In the Gender column, the values are derived as “Male” if the value in the Description field contains “Mr.” and “Female” for all the remaining values.

After extracting metadata fields, the data file can be seen as below:

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

In the above data, some of the values for “Age” column are present as NAs. These are the missing values in the “Age” data field. These missing values needs to be imputed using one of the imputing methods defined.

Implemented Mean data imputation method to impute the missing values in the Age column. This method takes the average or mean of all the values that are present in the Age column and replaces the missing values i.e., NAs with this mean value.

The data field “Age” before imputing with the mean can be shown as below:

```
> Data1$Age
  [1] 22 38 26 35 35 NA 54  2 27 14  4 58 20 39 14 55  2 NA 31 NA 35 34 15 28  8 38
 [27] NA 19 NA NA 40 NA NA 66 28 42 NA 21 18 14 40 27 NA  3 19 NA NA NA NA 18  7 21
 [53] 49 29 65 NA 21 28  5 11 22 38 45  4 NA NA 29 19 17 26 32 16 21 26 32 25 NA NA
 [79]  0 30 22 29 NA 28 17 33 16 NA 23 24 29 20 46 26 59 NA 71 23 34 34 28 NA 21 33
[105] 37 28 21 NA 38 NA 47 14 22 20 17 21 70 29 24  2 21 NA 32 32 54 12 NA 24 NA 45
[131] 33 20 47 29 25 23 19 37 16 24 NA 22 24 19 18 19 27  9 36 42 51 22 55 40 NA 51
[157] 16 30 NA NA 44 40 26 17  1  9 NA 45 NA 28 61  4  1 21 56 18 NA 50 30 36 NA NA
[183]  9  1  4 NA NA 45 40 36 32 19 19  3 44 58 NA 42 NA 24 28 NA 34 45 18  2 32 26
[209] 16 40 24 35 22 30 NA 31 27 42 32 30 16 27 51 NA 38 22 19 20 18 NA 35 29 59  5
[235] 24 NA 44  8 19 33 NA NA 29 22 30 44 25 24 37 54 NA 29 62 30 41 29 NA 30 35 50
[261] NA  3 52 40 NA 36 16 25 58 35 NA 25 41 37 NA 63 45 NA  7 35 65 28 16 19 NA 33
[287] 30 22 42 22 26 19 36 24 24 NA 23  2 NA 50 NA NA 19 NA NA  0 NA 17 30 30 24 18
[313] 26 28 43 26 24 54 31 40 22 27 30 22 NA 36 61 36 31 16 NA 45 38 16 NA NA 29 41
[339] 45 45  2 24 28 25 36 24 40 NA  3 42 23 NA 15 25 NA 28 22 38 NA NA 40 29 45 35
[365] NA 30 60 NA NA 24 25 18 19 22  3 NA 22 27 20 19 42  1 32 35 NA 18  1 36 NA 17
[391] 36 21 28 23 24 22 31 46 23 28 39 26 21 28 20 34 51  3 21 NA NA NA 33 NA 44 NA
[417] 34 18 30 10 NA 21 29 28 18 NA 28 19 NA 32 28 NA 42 17 50 14 21 24 64 31 45 20
[443] 25 28 NA  4 13 34  5 52 36 NA 30 49 NA 29 65 NA 50 NA 48 34 47 48 NA 38 NA 56
[469] NA  0 NA 38 33 23 22 NA 34 29 22  2  9 NA 50 63 25 NA 35 58 30  9 NA 21 55 71
[495] 21 NA 54 NA 25 24 17 21 NA 37 16 18 33 NA 28 26 29 NA 36 54 24 47 34 NA 36 32
[521] 30 22 NA 44 NA 40 50 NA 39 23  2 NA 17 NA 30  7 45 30 NA 22 36  9 11 32 50 64
[547] 19 NA 33  8 17 27 NA 22 22 62 48 NA 39 36 NA 40 28 NA NA 24 19 29 NA 32 62 53
[573] 36 NA 16 19 34 39 NA 32 25 39 54 36 NA 18 47 60 22 NA 35 52 47 NA 37 36 NA 49
[599] NA 49 24 NA NA 44 35 36 30 27 22 40 39 NA NA NA 35 24 34 26  4 26 27 42 20 21
[625] 21 61 57 21 26 NA 80 51 32 NA  9 28 32 31 41 NA 20 24  2 NA  0 48 19 56 NA 23
[651] NA 18 21 NA 18 24 NA 32 23 58 50 40 47 36 20 32 25 NA 43 NA 40 31 70 31 NA 18
[677] 24 18 43 36 NA 27 20 14 60 25 14 19 18 15 31  4 NA 25 60 52 44 NA 49 42 18 35
[703] 18 25 26 39 45 42 22 NA 24 NA 48 29 52 19 38 27 NA 33  6 17 34 50 27 20 30 NA
[729] 25 25 29 11 NA 23 23 28 48 35 NA NA NA 36 21 24 31 70 16 30 19 31  4  6 33 23
[755] 48  0 28 18 34 33 NA 41 20 36 16 51 NA 30 NA 32 24 48 57 NA 54 18 NA  5 NA 43
[781] 13 17 29 NA 25 25 18  8  1 46 NA 16 NA NA 25 39 49 31 30 30 34 31 11  0 27 31
[807] 39 18 39 33 26 39 35  6 30 NA 23 31 43 10 52 27 38 27  2 NA NA  1 NA 62 15  0
[833] NA 23 18 39 21 NA 32 NA 20 16 30 34 17 42 NA 35 28 NA  4 74  9 16 44 18 45 51
[859] 24 NA 41 21 48 NA 24 42 27 31 NA  4 26 47 33 47 28 15 20 19 NA 56 25 33 22 28
[885] 25 39 27 19 NA 26 32
```

These NA values are replaced with mean of the remaining values by mean imputation method using R script.

The mean obtained from the remaining values is **23.78339**. As this value cannot be taken as the age of a person, rounded of the value to the nearest integer using **Floor** function in R. Also, removed the Description column as the data in the column is split and added as new columns to each data record.

After imputing the mean values in the Age column, the data is ready to be used for further analysis. The data that is ready for analysis is as shown below:

CUSTOMERID	LastName	FirstName	Gender	Age	SUCCESS	SEATCLASS	GUESTS
1	Braund	Mr. Owen Harris	Male	22	0	3	1
2	Cumings	Mrs. John Bradley (Florence Briggs Thayer)	Female	38	1	1	1
3	Heikkinen	Miss. Laina	Female	26	1	3	0
4	Futrelle	Mrs. Jacques Heath (Lily May Peel)	Female	35	1	1	1
5	Allen	Mr. William Henry	Male	35	0	3	0
6	Moran	Mr. James	Male	23	0	3	0
7	McCarthy	Mr. Timothy J	Male	54	0	1	0
8	Palsson	Master. Gosta Leonard	Female	2	0	3	3
9	Johnson	Mrs. Oscar W (Elisabeth Vilhelmina Berg)	Female	27	1	3	0
10	Nasser	Mrs. Nicholas (Adele Achem)	Female	14	1	2	1
11	Sandstrom	Miss. Marguerite Rut	Female	4	1	3	1
12	Bonnell	Miss. Elizabeth	Female	58	1	1	0
13	Saundercock	Mr. William Henry	Male	20	0	3	0
14	Andersson	Mr. Anders Johan	Male	39	0	3	1
15	Vestrom	Miss. Hulda Amanda Adolfina	Female	14	0	3	0
16	Hewlett	Mrs. (Mary D Kingcome)	Female	55	1	2	0
17	Rice	Master. Eugene	Female	2	0	3	4
18	Williams	Mr. Charles Eugene	Male	23	1	2	0
19	Vander Planke	Mrs. Julius (Emelia Maria Vandemoortele)	Female	31	0	3	1
20	Masselmani	Mrs. Fatima	Female	23	1	3	0

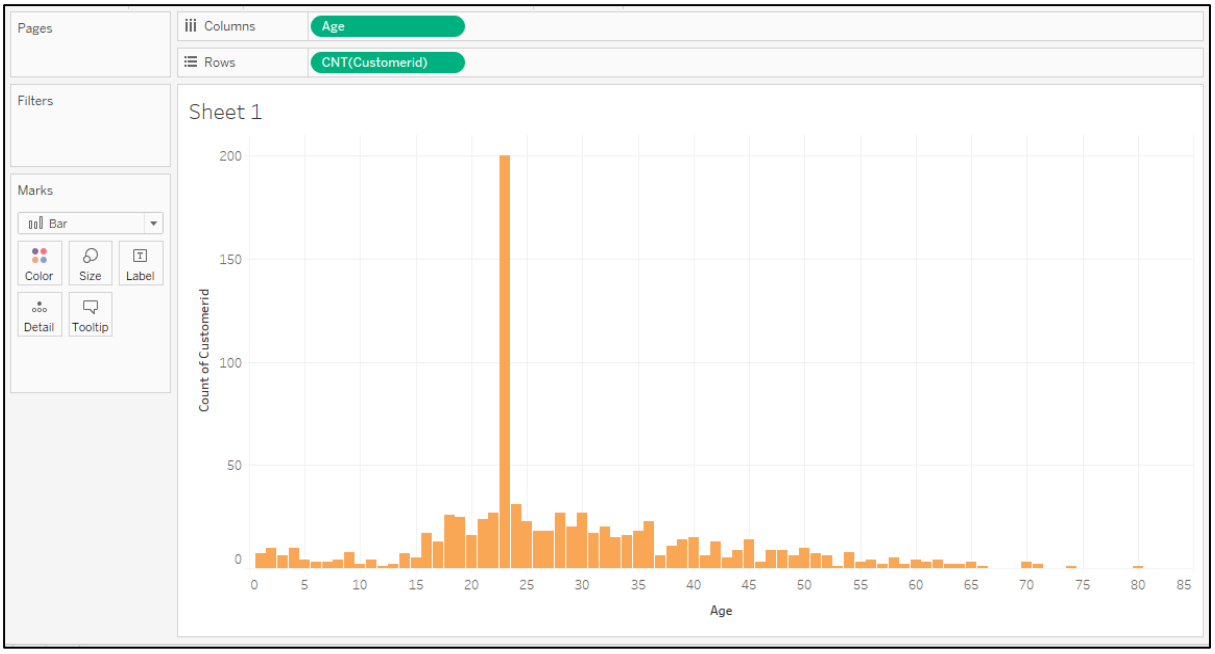
This data file in CSV is ready to be analysed further for next steps i.e., Metadata exploration to observe the patterns of distribution between each of the metadata.

iii. Metadata Exploration

This milestone is achieved by observing different patterns of distribution between each of the metadata compared with Customer ID field using visualization tool Tableau.

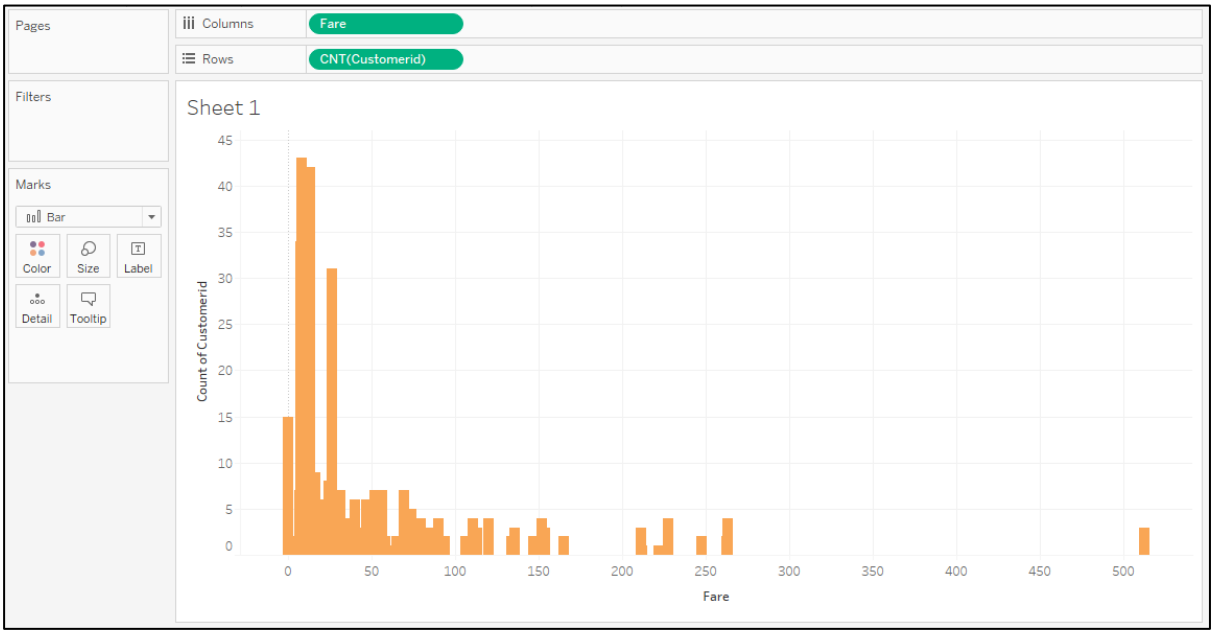
Plotted graphs for these comparisons in Tableau. Below are the graphs plotted:

a. Number of Customers Vs Age



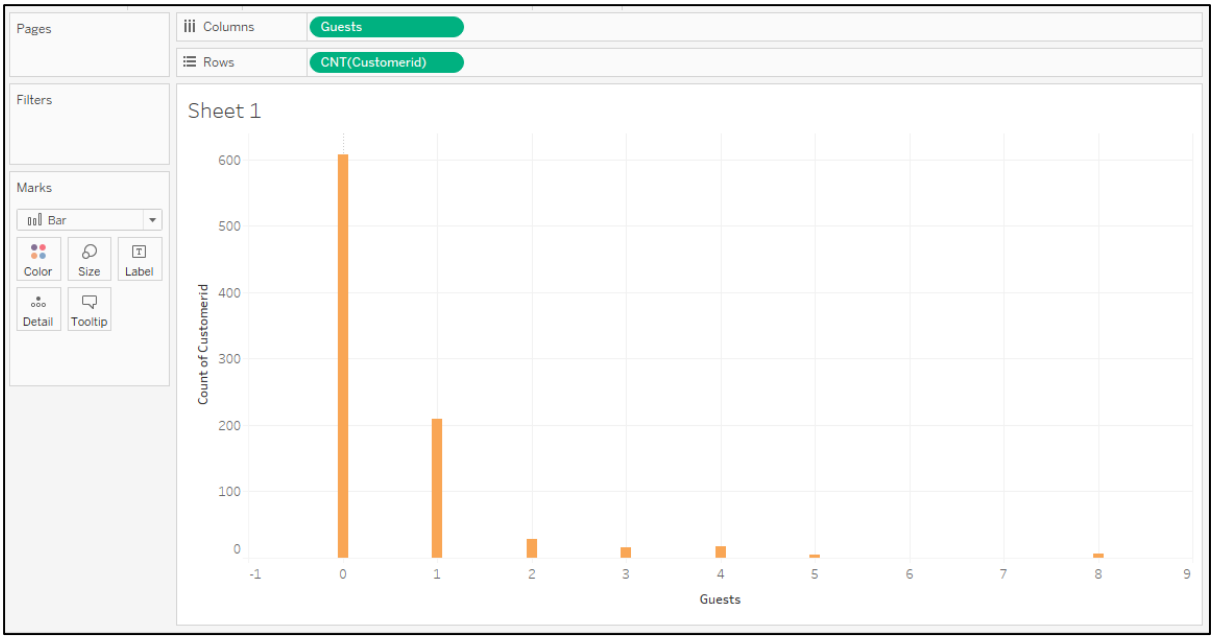
From the above plot, we can observe that customers aging 23 are 200 i.e., more than any other age group in the dataset. This means that customers around age 23 are flying more.

b. Number of Customers Vs Fare



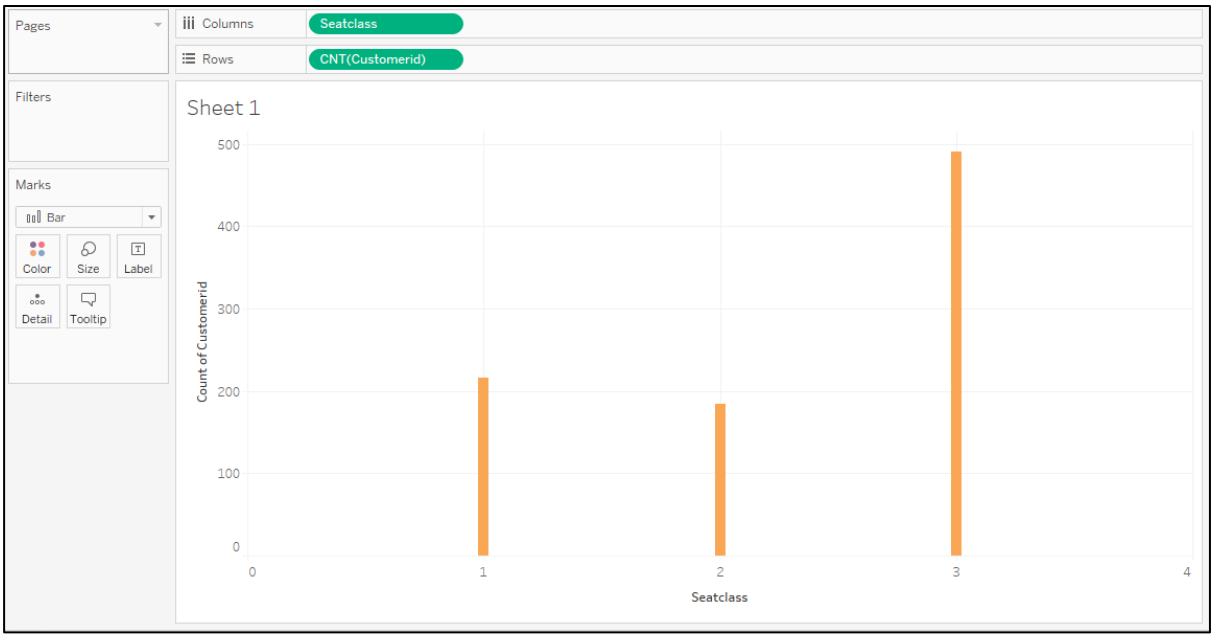
From the above plot, we can observe that Customers are more within the fare range of 7 to 26. This means that more number of customers are willing to fly when the fares are low. As the fares kept on increasing from 26, the number of customers flying went on decreasing.

c. Number of Customers Vs Guests



From the above plot, we can observe that number of Customers is 608 when the accompanying Guests is 0. This means that the customers are willing to travel alone or more number of Customers will travel when there are less guests or no guests accompanying them.

d. Number of Customers Vs Seat Class

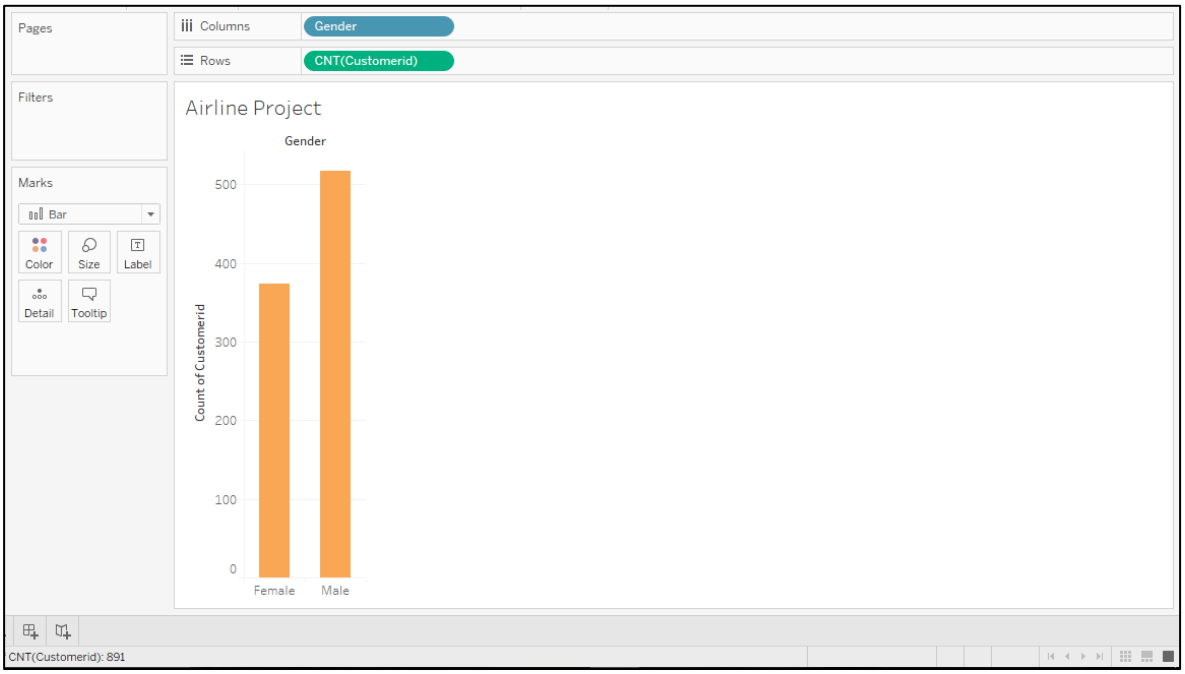


From the above plot, we can observe that number of customers are 216 for seat class 1, 184 for seat class 2 and 491 for seat class 3. This means that more number of customers are flying in seat class 3.

e. Number of Customers Vs Success

From the above plot, we can observe that Number of customers is 549 when the success is 0 and the number of customers is 342 when the success is 1. This means more number of customers are not flying.

f. Number of Customers Vs Gender



From the above plot, we can observe that the number of customers are 517 when they are male and the number is 374 when they are female. This clearly means that more number of Males are flying than Females.

All the above plots give information to help in further analysis in developing a prediction model and drawing insights to help recommend the advertising team to develop customer demographic specific models.

iv. Attribute Preparation and Engineering for preparing for Mining Algorithm

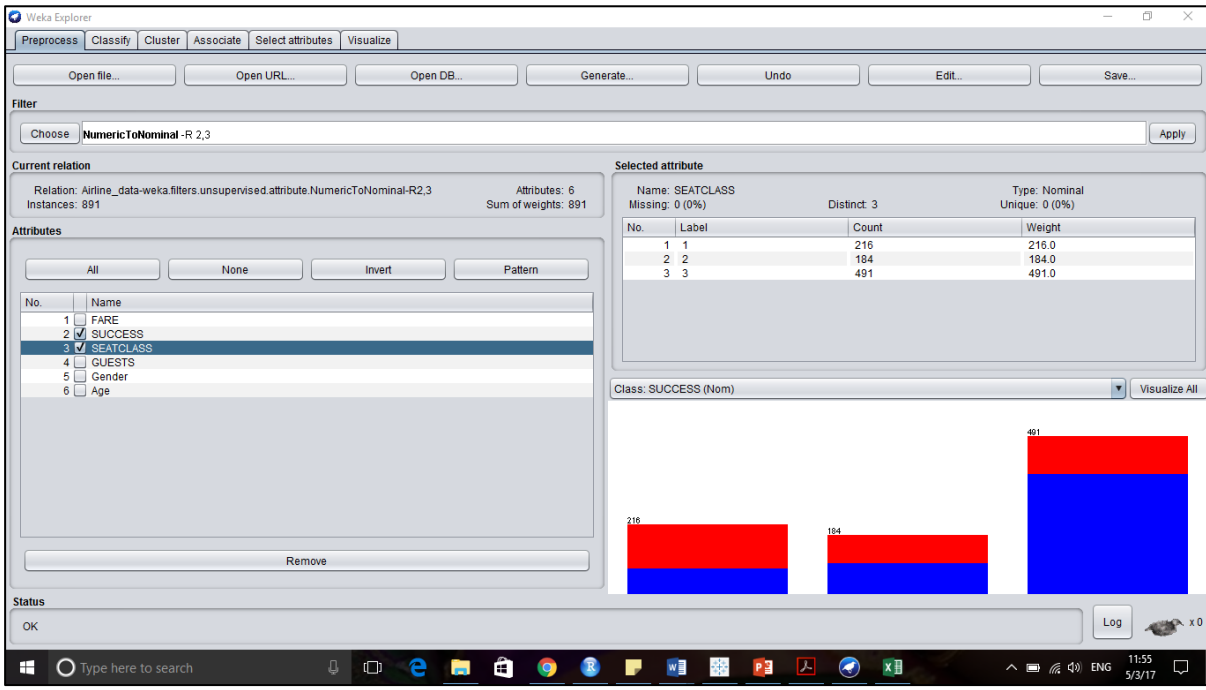
There are two steps involved in this milestone. The first step is to load the CSV file obtained in the above steps to Weka and change the type of the attribute to Nominal to develop a model. And the second step is to identify top 2 attributes from the obtained data to develop the model.

a. Import data file into Weka

The CSV file that is obtained in the above steps is to be loaded into Weka for further analysis. Once the data file is loaded, the attribute should be changed to right type in order to process the data further and develop the model.

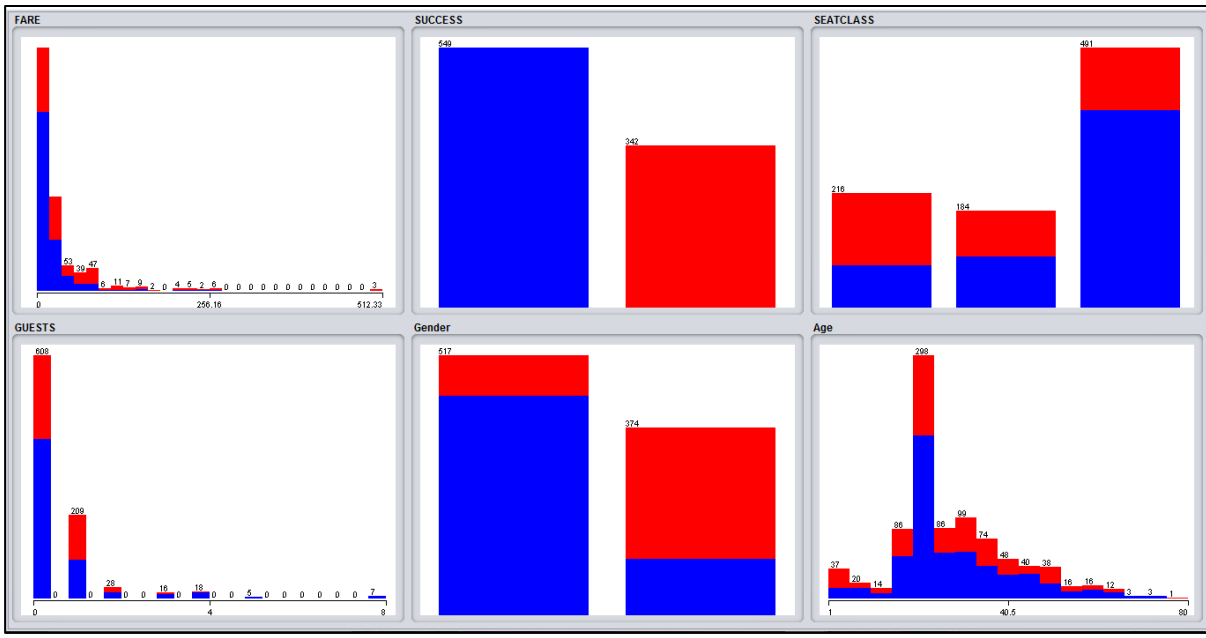
Data is imported into Weka tool in the “Preprocess” tab by clicking on “Open file” button and giving the path where the data file is located.

After loading the data into the tool, attribute type is changed to “NumericToNominal” using Filter option. All the attributes in the data file will be loaded into Weka for analysis. Plotting Success vs Seat Class as attributes, obtained the below plots:



In the above plot, blue denotes success 0 and red denotes success 1 and plot shows success distribution between each class i.e., class 1, 2 and 3.

Similarly, plotting graphs for the others attributes by clicking “Visualize All” button, the graph obtained is as below:



The above plots show the distribution between all the different attributes that are extracted from the data file.

b. Attribute selection

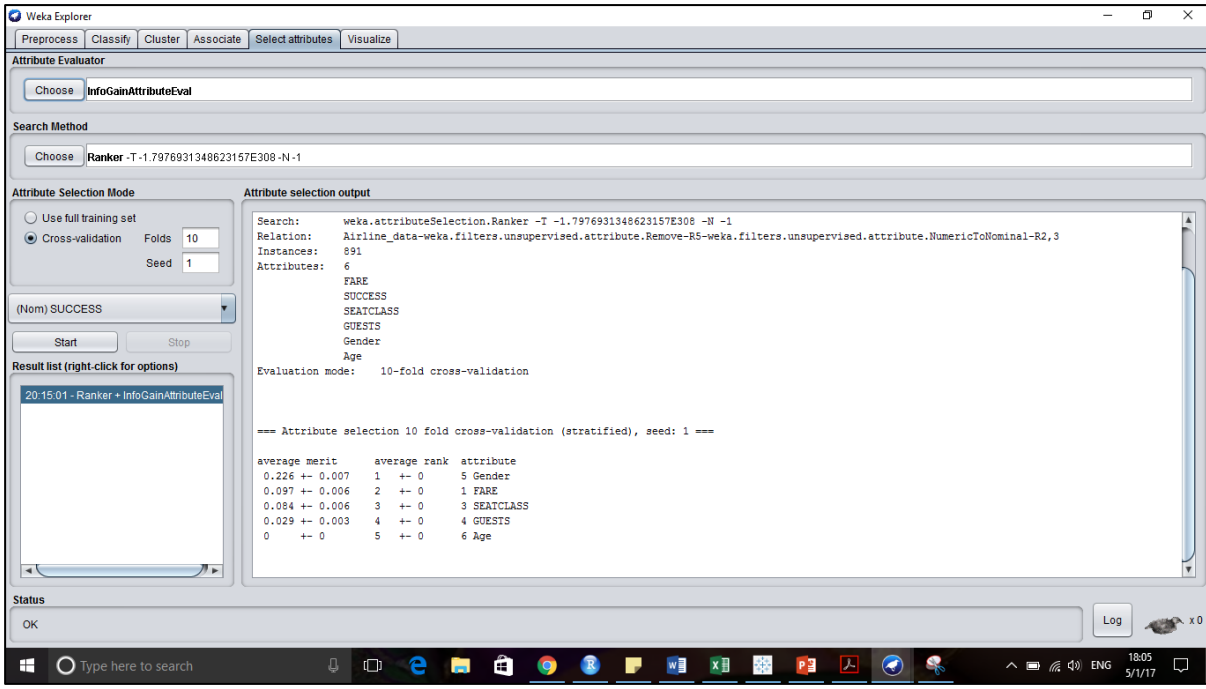
The objective of this step is to find top 2 attributes with 10-fold cross validation to develop the model. For this, Attribute evaluator with ranker method is run on the data in Select Attribute tab of Weka. Chose “InfoGainAttributeEval” as attribute evaluator with “Ranker” as search method and Attribute selection mode as Cross validation with number of folds as 10, which gives the ranks for attribute selection.

After running this method, ranks of the attributes are generated as below:

From the below results, ranks of each attribute are clearly seen. The top 2 attributes being **Gender** and **Fare**.

```
average merit      average rank  attribute
 0.226 +- 0.007    1    +- 0      5 Gender
 0.097 +- 0.006    2    +- 0      1 FARE
 0.084 +- 0.006    3    +- 0      3 SEATCLASS
 0.029 +- 0.003    4    +- 0      4 GUESTS
 0      +- 0       5    +- 0      6 Age
```

Results are shown below:



v. Prediction Modelling and Visualization

The objective of this milestone is to design a classification model using Decision Tree and Random Forest algorithms and also to generate ROC curves for these models with 10-fold cross validation.

1. Classification model

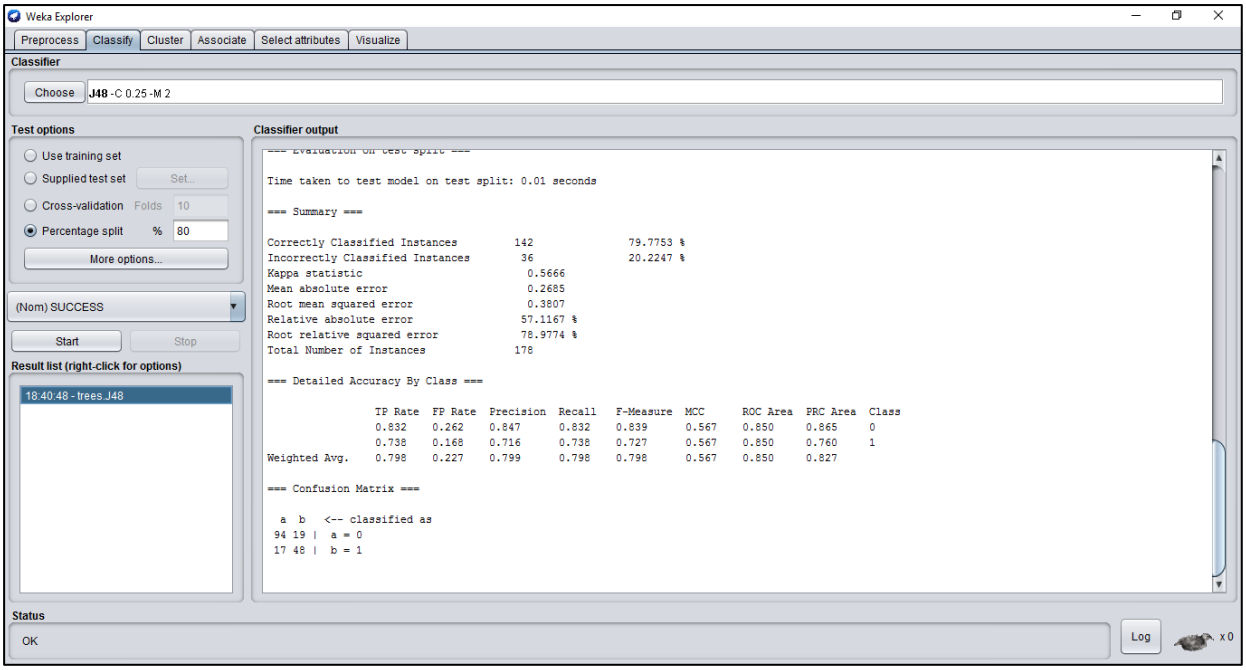
a. Decision Tree:

For this classification model, divided the data into Training data and Test data by using default classification feature “Percentage Split” of Weka. Divided the data into 80% Training data and 20% Test data. After running the model Decision Tree **J48** on 80% Training data, below are the results obtained:

Correctly Classified Instances	142	79.7753 %
Incorrectly Classified Instances	36	20.2247 %
Kappa statistic	0.5666	
Mean absolute error	0.2685	
Root mean squared error	0.3807	
Relative absolute error	57.1167 %	
Root relative squared error	78.9774 %	
Total Number of Instances	178	

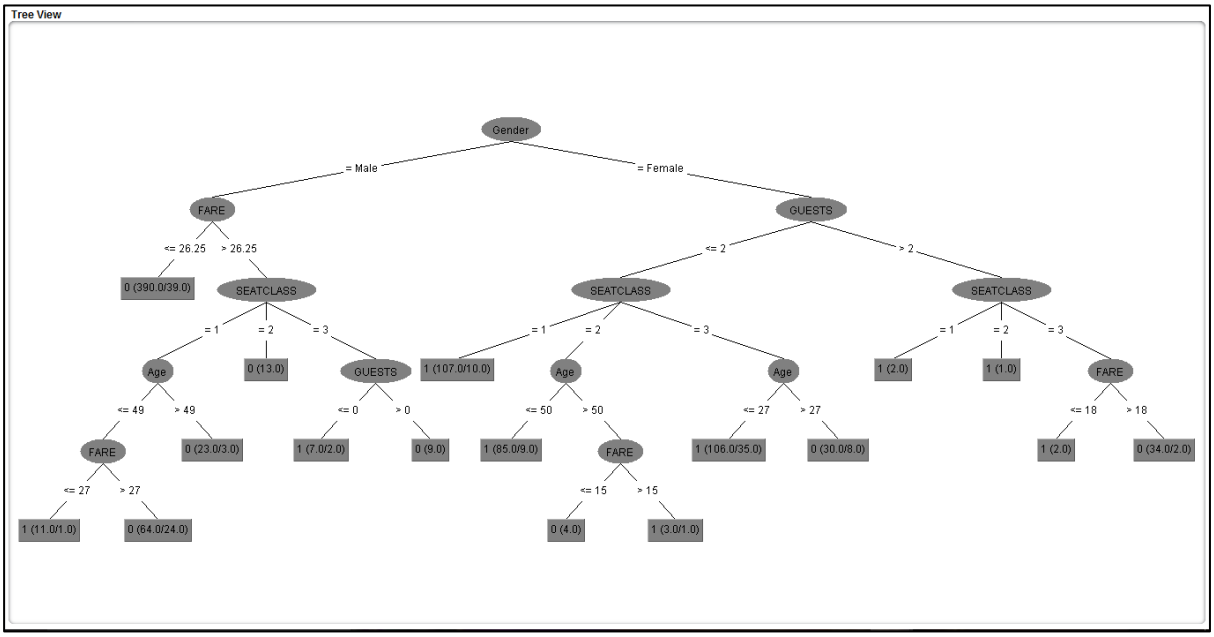
=== Confusion Matrix ===

a	**b**	**<-- classified as**
94	**19**	**| a = 0**
17	**48**	**| b = 1**



From the above results, we can observe that the accuracy of this model as **79.77%** with 142 correctly classified instances.

Decision tree plotted for the model,



Decision tree is been built with parent node as **Gender.**

b. **Random Forest:**

In the similar way, as for the Decision Tree, data is been divided as training and test data for this model as well. After running the model on 80% training data, below are the results:

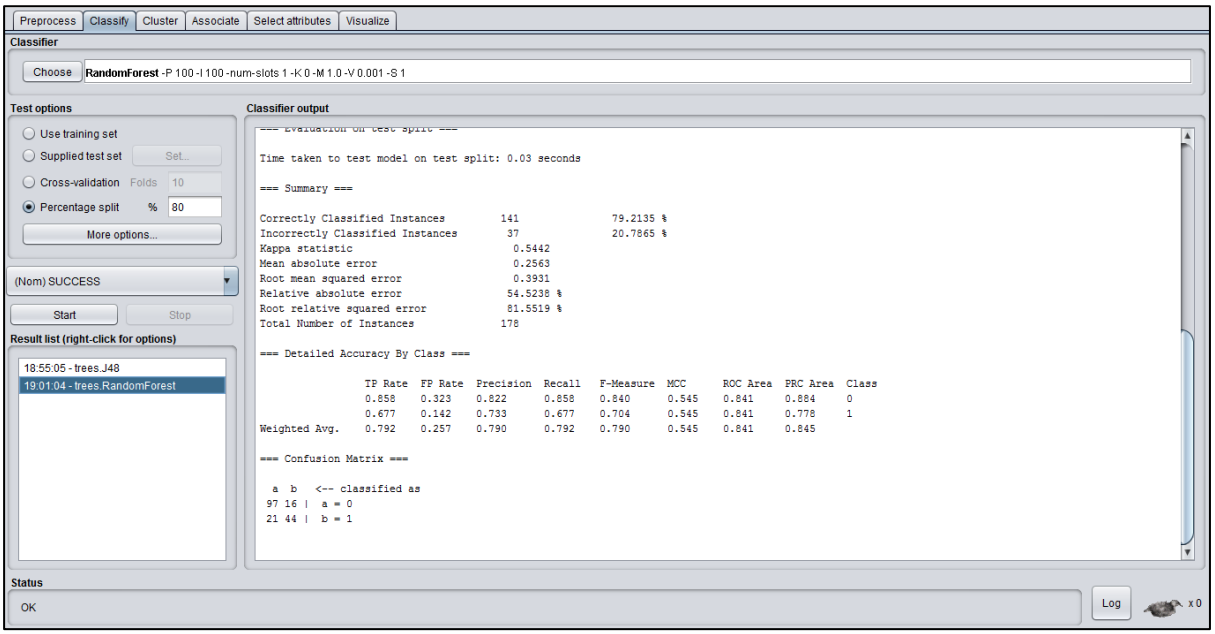
Correctly Classified Instances	141	79.2135 %
Incorrectly Classified Instances	37	20.7865 %
Kappa statistic	0.5442	
Mean absolute error	0.2563	
Root mean squared error	0.3931	
Relative absolute error	54.5238 %	
Root relative squared error	81.5519 %	
Total Number of Instances	178	

=== Confusion Matrix ===

a b <-- classified as

97 16 | a = 0

21 44 | b = 1



From the above results, we can observe that the accuracy of this model as **79.21%** with 141 correctly classified instances.

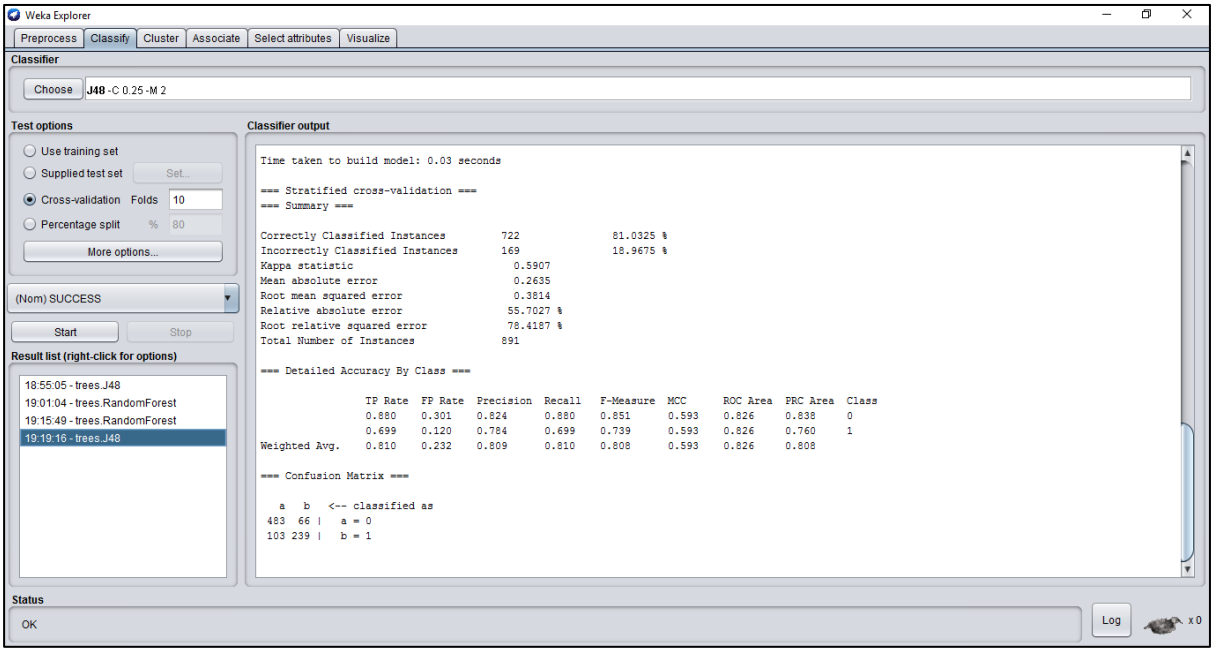
From both the models we can understand that Decision tree algorithm with high accuracy is the better model.

2. ROC Curves:

A Receiver Operating Characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. Plotting ROC curve using Decision Tree and Random Forest algorithms with 10-fold cross validation method, obtained the below plots:

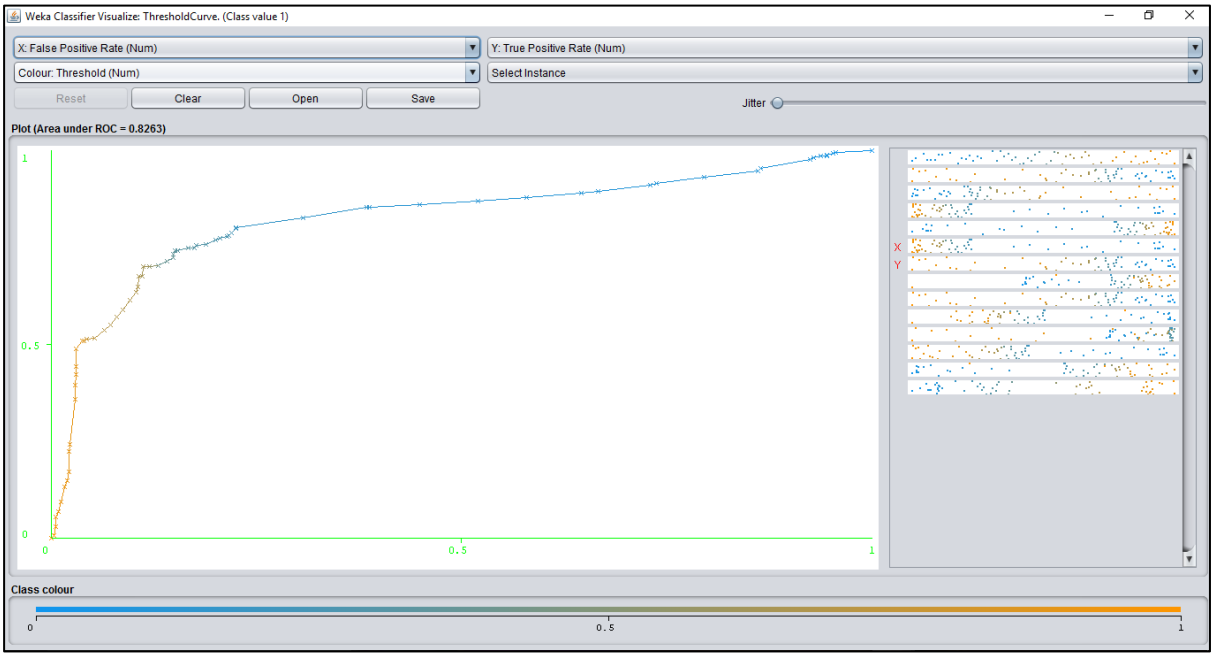
a. ROC curve – Decision Tree (J48)

Decision Tree classification model is run with 10-fold cross validation, results are obtained as below,



Accuracy is **81.03%** with 722 correctly classified instances.

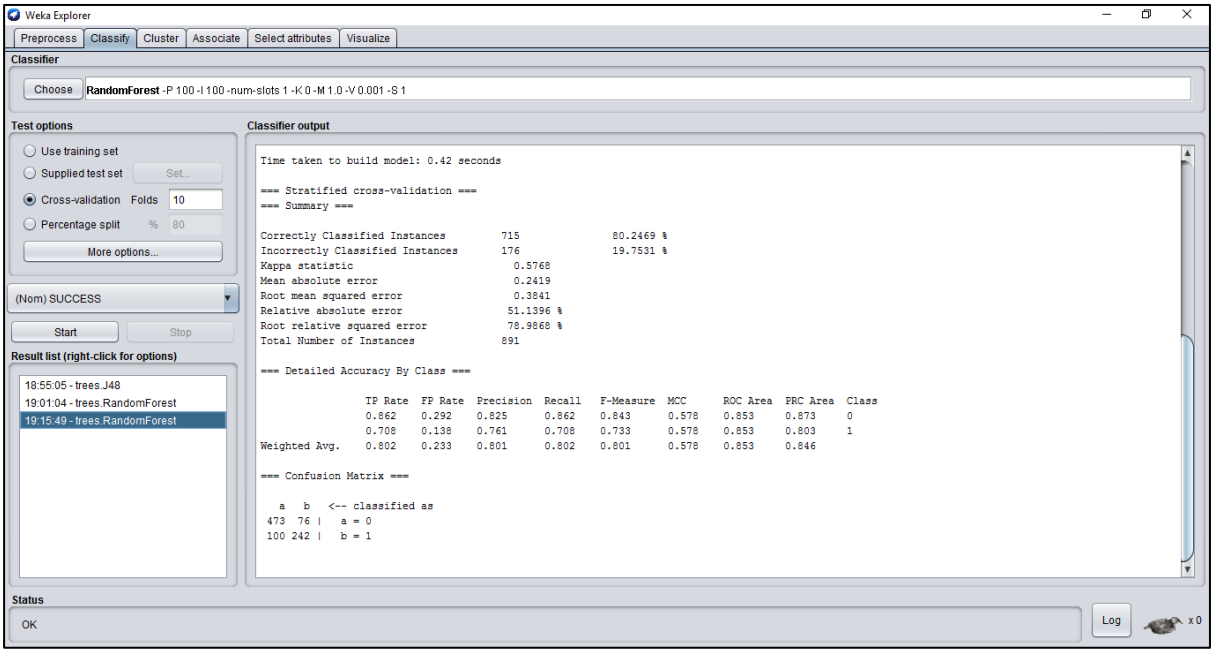
ROC curve for threshold 1 is shown below,



As seen from the plot, the area under curve obtained is **0.8263**, which means that this is a good model. As also seen from the plot, for a good model, the curve should be bending towards Y-axis which also suggests this as a good model.

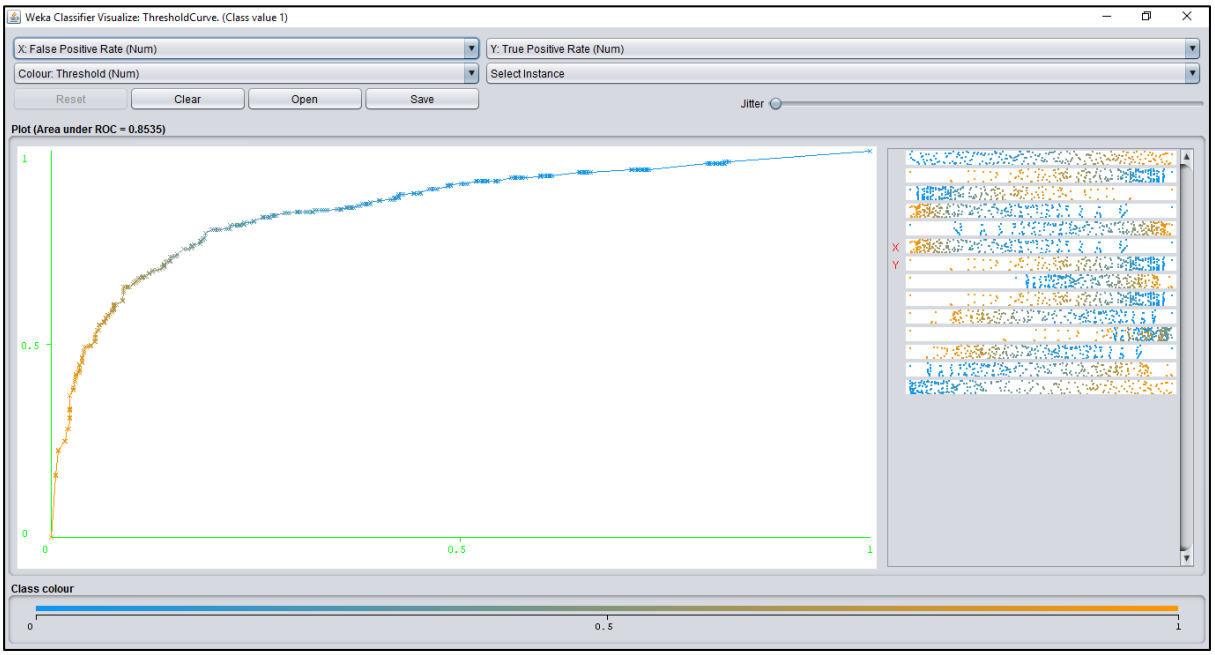
b. ROC Curve – Random Forest

Random Forest classification model is run with 10-fold cross validation, results are obtained as below,



Accuracy of this model as seen from the above results is **80.24%** with 715 correctly classified instances.

ROC curve for this model with threshold value 1 is shown below,



As observed from the above plot, the area under curve is **0.8535**. Though the area under curve is more than the area for Decision Tree model, as the accuracy is more for Decision Tree model and also the curve for Random Forest method is smoother than Decision tree model curve, Decision Tree method is more better model than Random Forest model.

Insights:

As observed from the above results from all the milestones, we can draw few insights and recommend these factors to advertising team of the company to design demographic specific plans and thereby increase their sales.

1. People aged around 23 travel more
2. Success rate is high for males
3. Top 2 attributes to predict the model are Gender and Fare.

If the team can work on these factors, there will be more chances that their sales can pick up and get high success rate with more number of customers flying.