

AIRLINE PROJECT REPORT

BY SAMEER KUMAR



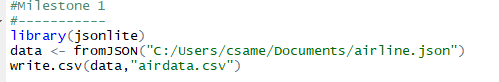
May 1, 2018

GEORGE MASON UNIVERSITY

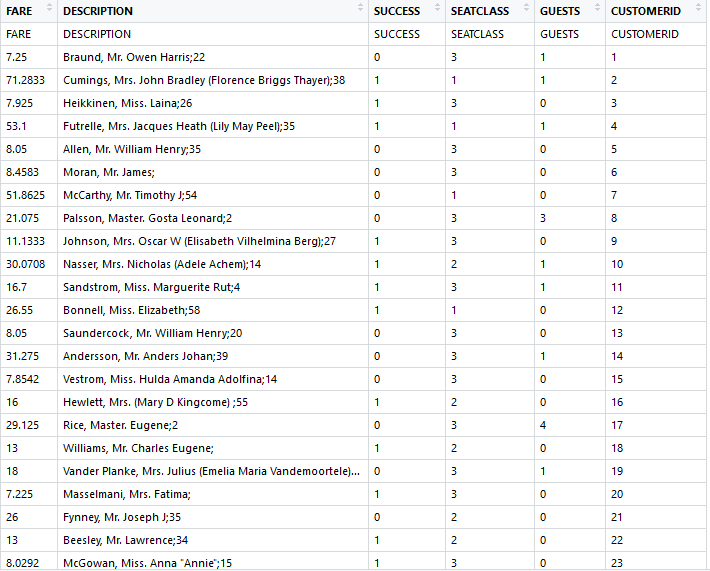
G NO. – G01099808

**MILESTONE #1**

**DATA ACQUISITION & CONVERSION**



#Converted CSV File & Resulting Data in R.

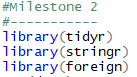


* The dataset is not clean and has sparse data values.
* The attributes are also not atomic.

**MILESTONE #2**

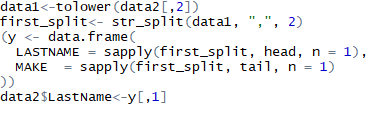
**METADATA EXTRACTION & IMPUTATION**

Libraries used for metadata extraction & imputation:

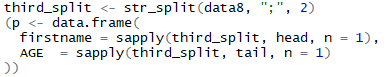


Few extraction & imputation commands:

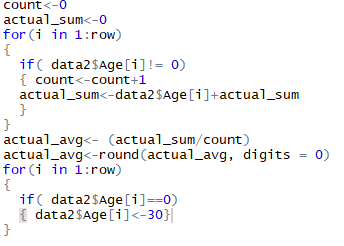
1. Last name Extraction



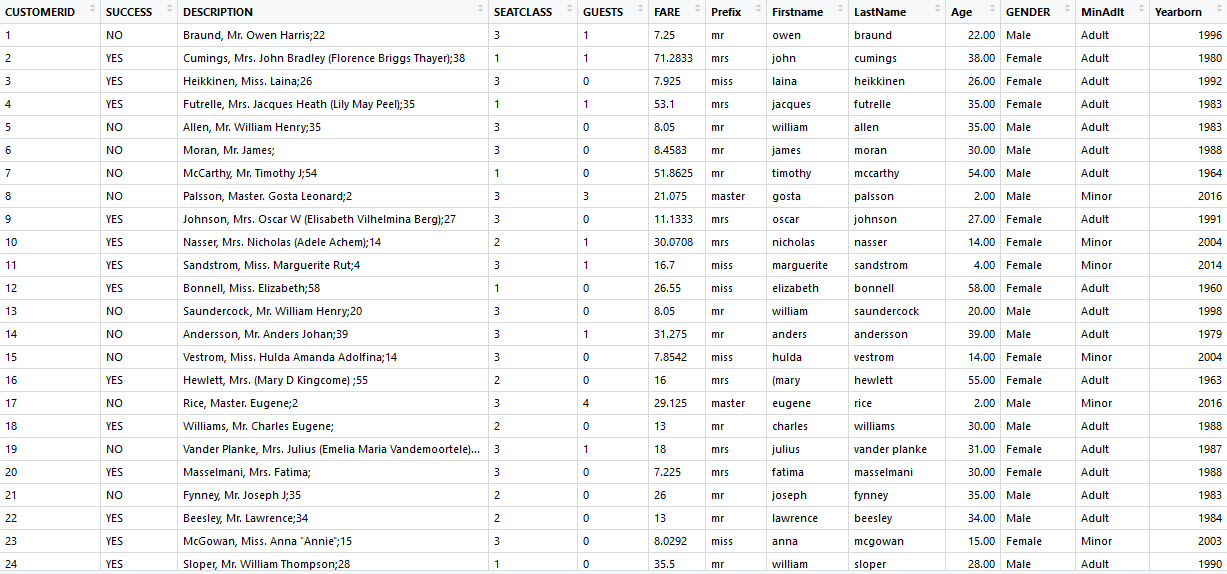
1. Age Extraction



1. Imputation of Missing variable Age



Final Extracted and Imputed Data set:

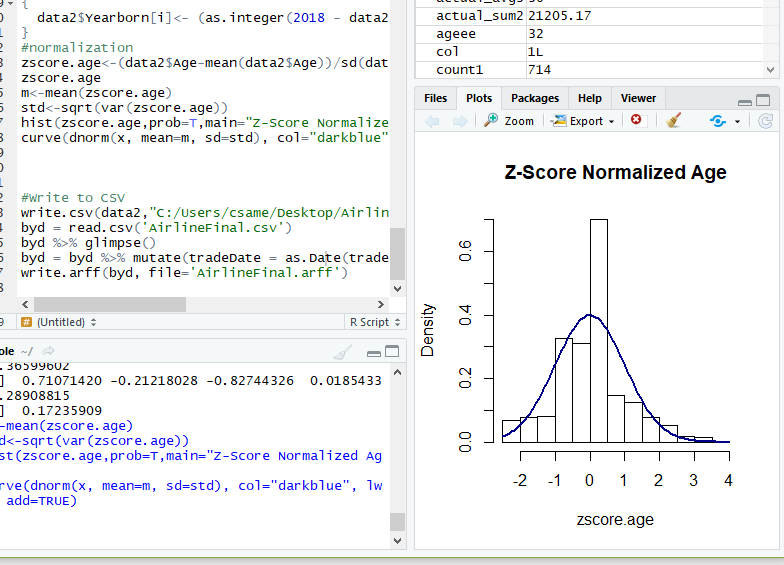


List of Extracted Metadata:

1. Prefix
2. First Name
3. Last Name
4. Age
5. Gender(M/F)
6. Minor/Adult Category (18+ Adult)
7. Year born (Age subtracted from today’s date)

#Normalization of Age & Fare

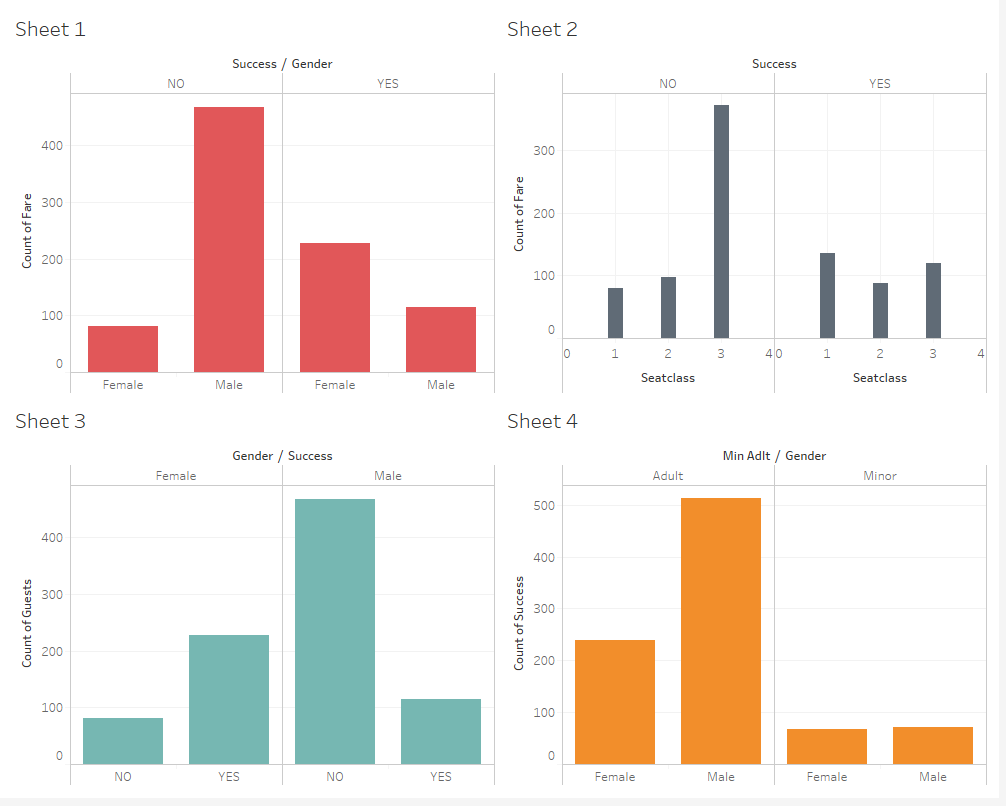
* Age group was normalized using the Z-score method as it is best suited for attributes having more null or zero values



* Fare column doesn’t need to be normalized using the Z-score Normalization method for optimum results as it is a non-numeric column and normalization is not helpful for categorical values.

**MILESTONE #3**

**METADATA EXPLORATION**

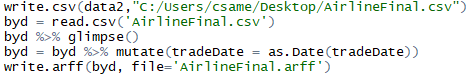


* Figure 1: The graphs depicts the relation between success rate and gender category with respect to Fare count. The positive success rate is higher for females than males. On the other hand, the negative success rate is higher for males than females.
* Figure 2: The graphs depicts relation of success rate with respect to seat class and Fare. As describes in the figure, the negative success rate has higher latency for 3rd seat group that means more number of people cancelled the airline ticket who were having seat class 3. The positive success rate is almost similar for all the class seats but precisely seat class 1 has higher frequency.
* Figure 3: The graph depicts the relation between gender and success along with number of guests. It shows that higher positive success rate in females than males with respect to number of guests and higher frequency of negative success rate for males than females.
* Figure 4: The figure briefs about the relationship between different age groups i.e. Adult & Minor for success rate with respect to gender. The success rate is higher for Adult Male than females and the rate is similar for minors either male or female.

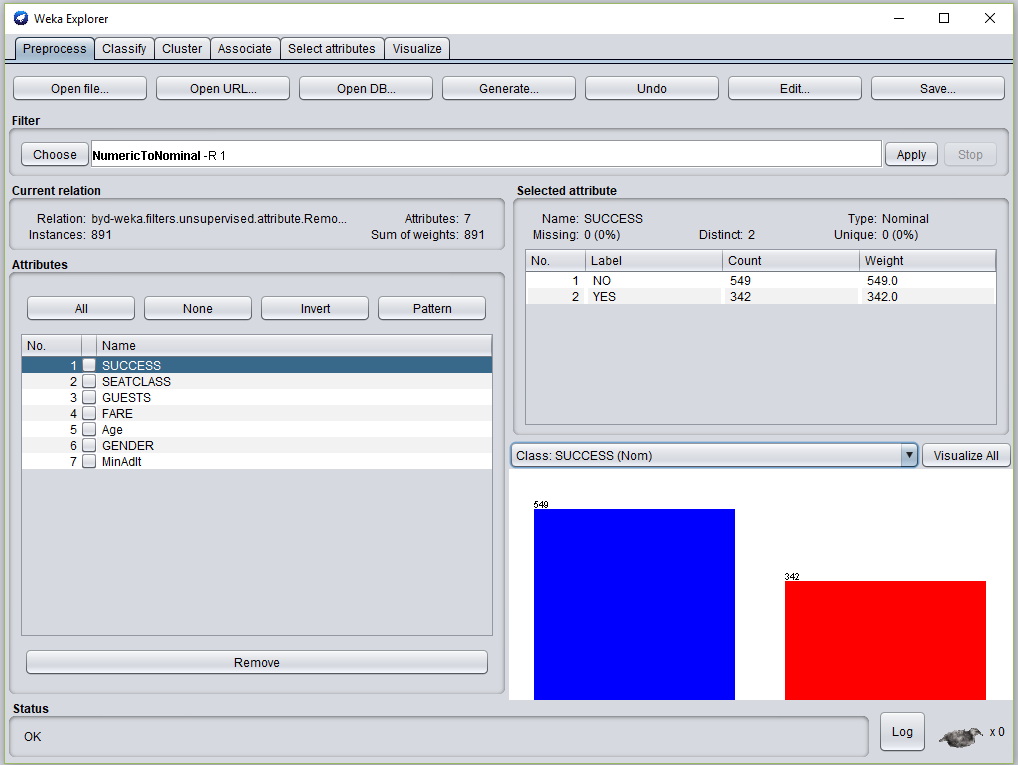
**MILESTONE #4**

**ATTRIBUTE PREPARATION & ENGNEERING FOR PREPARING FOR MINING ALGORITHM**

1. Conversion of CSV file to ARFF using R-Studio.

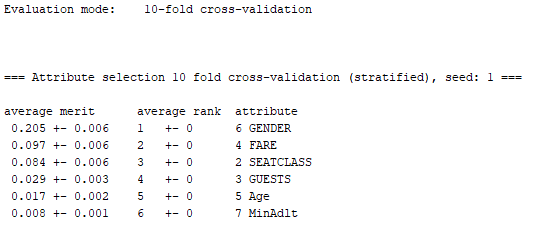


1. Using Weka to perform attribute selection on the modified data set.



In the above figure, we performed the attribute selection algorithm and applied numeric to nominal filter on the success attribute. The blue color indicates positive success rate for the data set and red color shows the vice versa. The positive success has a higher frequency for the given data set.

The top two attributes are calculated using the 10-fold cross validation as follows:



Top attribute <- Gender

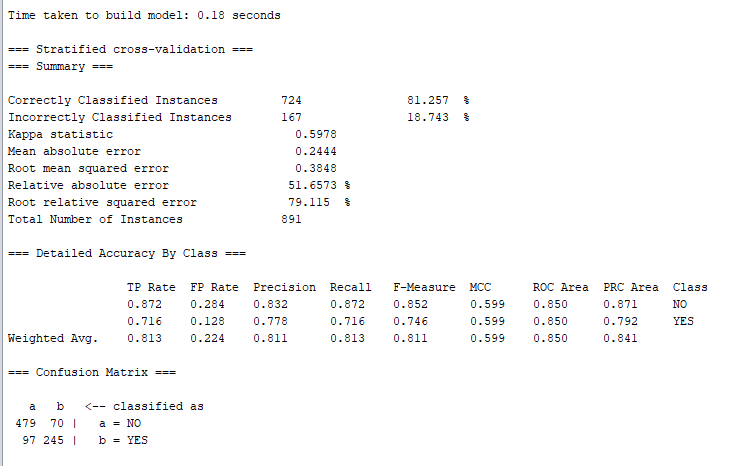
Second attribute <- Fare

The above two attributes have the most affect on the success rate as compared to other attributes. According to the data set visualization graphs, male customers are not likely to succeed as compared to female customers.

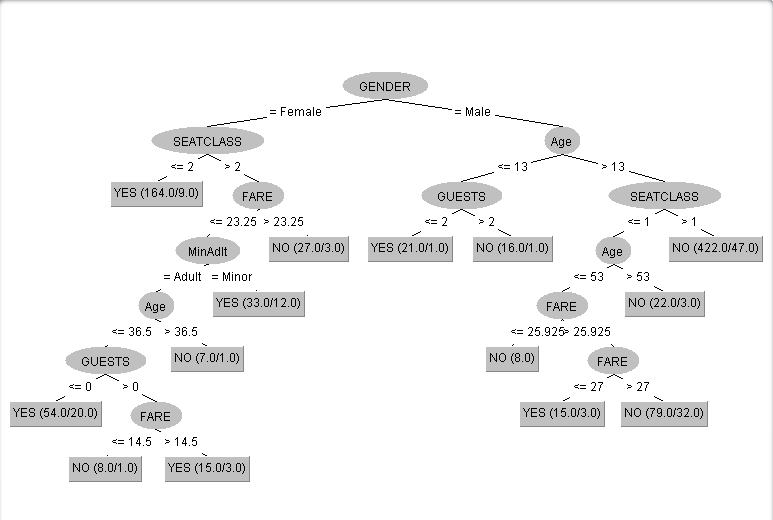
**MILESTONE #5**

**PREDICTION MODELLING AND VITUALIZATION**

1. Classification matrix with accuracy & error percentage. (Confusion Matrix) (J48)



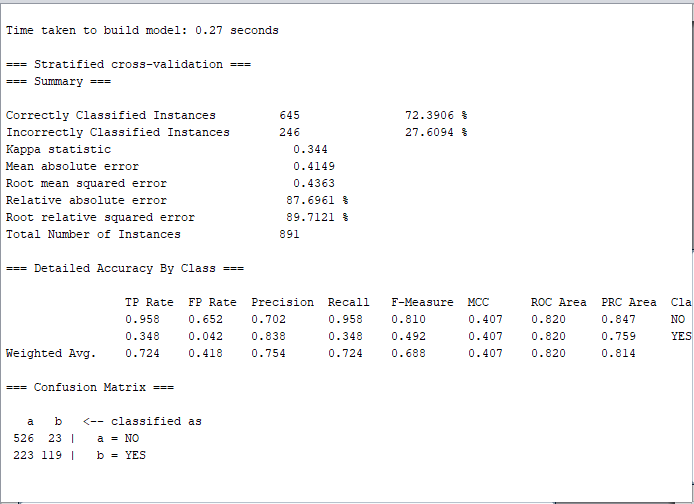
1. Decision Tree



Approx. Accuracy for J48: 81%

Root mean square error for J48: 0.38

Classification model for Random Forest:



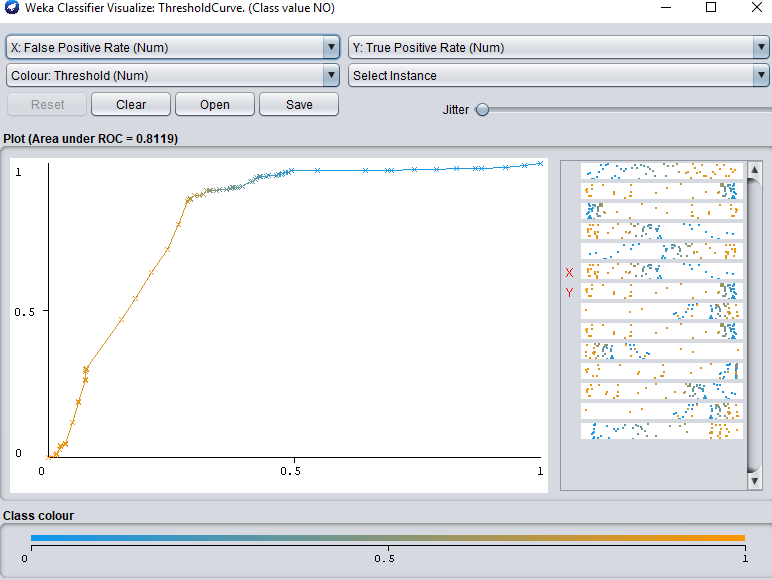
Approx. Accuracy for Random Forest: 72%

Root mean square error for Random Forest: 27

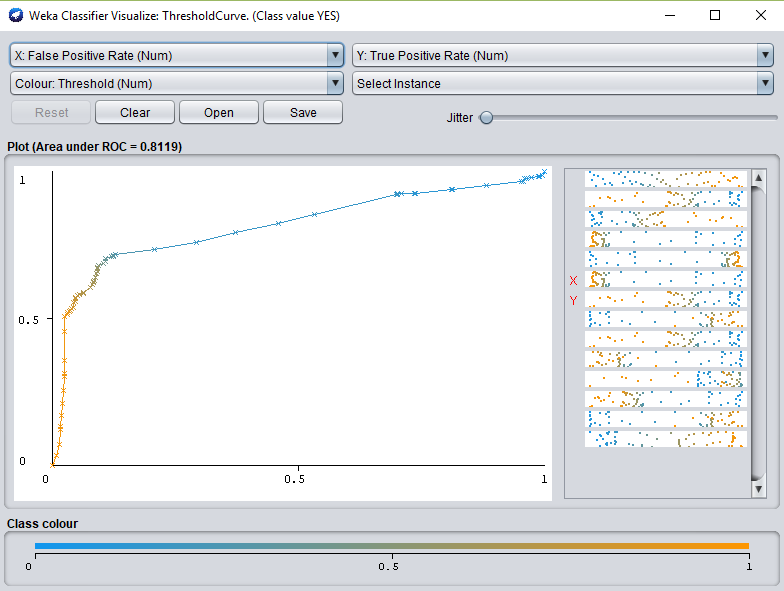
Therefore, the better of the two classification model build is J48 with higher accuracy and less error possibility.

1. ROC Curve for Random Forest

Class Value: NO

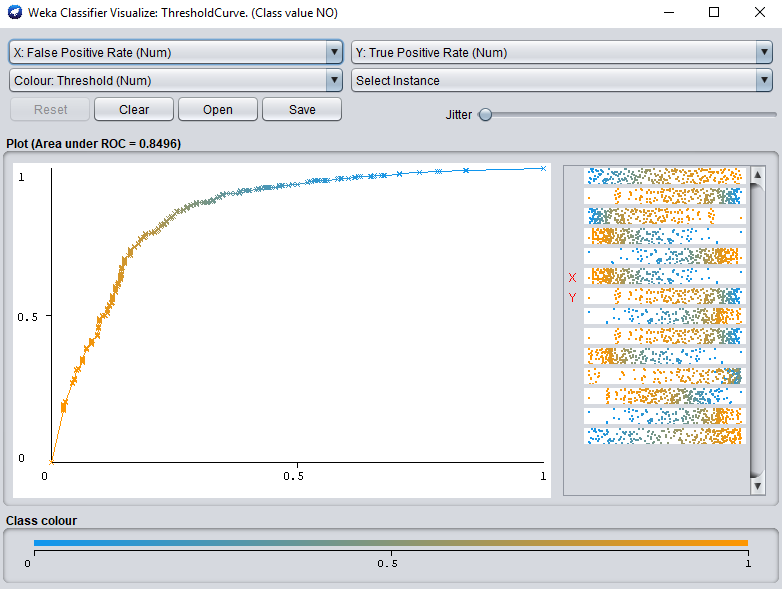


Class Value: YES

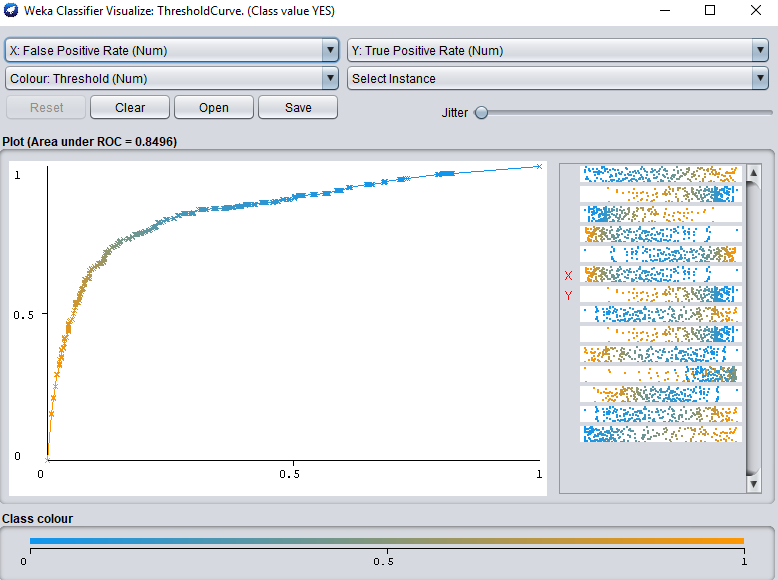


1. ROC Curve for J48

Class Value: NO



Class Value: YES



The larger the area under the ROC curve is the more accurate the model is considered.

The graph is plotted between TPR & FPR for different class values (yes/no). J48 and random forest are two different classification trees with different ROC curves.

#Area Calculation for both models:

🡪As given in the ROC curve for J48, the area under the curve for threshold value 1 is 0.819, which can be considered as a large portion of the graph and inclined towards the y axis. The more the area under the graph, the better the model is being considered.

🡪For ROC curve in random forest modelling, the value under the graph for maximum threshold 1, the area value is 0.849 which can be said to be an excellent model.

**Goal of the project** 🡪 Recommendations for Airline company to reduce cancellations:

1. Target the economy class 3 passengers thereby giving promotional offers to them.
2. The male passengers are highlighting factor and the company can offer some extra free meals or reduce the fair price for these customers so that they do not cancel the flight.
3. Reduce the price for all seat class so that people do not care about the refund and do not cancel the flight ticket.
4. Airline company should offer free guest service to those having higher number of guest with them.
5. They should also reduce fare for male passengers with more number of guests and target these guests for offers and promotions.
6. Lifetime promotional offer of carrying a guest for free if the number of total passengers is more than 5.
7. Extra baggage allowance must be given so that students and needy people do not change the airline and cancel the flight at the last moment.