Summary Report

Sai Gopi Krishna Govindarajula

Introduction : The data set with which we are working over here is a classification data set. It is used to classify between economy prices and premium prices of tickets in a flight.

Several adjourning factors are also provided in the data set in order for us to understand what actually is going on in the data set.

From the analysis we see many inferences. I will conclude the inferences that I have made in the form of points for better understanding of the overall audience. These are done by executing code snippets one after the other and giving an overall view on the entire case :

1. Most number of flights are by British airlines.
2. International flights are more than domestic flights.
3. Low range Economy tickets are more frequent.
4. There is a positive correlation between cost of both the type of tickets and the flight duration.
5. Positive correlations are seen in case of Price\_Economy between quality and international flight and width\_economy and flight duration.
6. Positive correlations are seen between, quality and pitch and width premium, width\_premium and international, pitch\_premium and international.
7. Positive correlation is seen between prices of each class and pitch and width.
8. Most important predictor for price Economy is Flight duration and price relative.
9. Most important predictor for price Premium is Flight duration and price relative.
10. the t value of Pitch\_economy and quality is positive indicating that these predictors are associated with

Price\_economy. A larger t-value indicates that that it is less likely that the coefficient is not equal to zero purely by chance.

Again, as the p-value for Flight\_Duration and Price\_Relative is less than 0.05 they are both statistically significant in the multiple linear regression model for Price\_Economy response variable.

The model’s, p-value: < 2.2e-16 is also lower than the statistical significance level of 0.05, this indicates that we can safely reject the null hypothesis that the value for the coefficient is zero

(or in other words, the predictor variable has no explanatory relationship with the response variable).

The model has a F Statistic of 90, which is considerably high

1. The r squared and the adjusted r squared is relatively high signifying a good model, suggesting that pitch, width, relative price, quality and flight duration add to the price of a ticket.
2. From the VIF plot we see that Flight Duration and Price Relative are most important factors in predicitng Price Economy.
3. We get an accuracy of 0.563 using the Random Forest Algorithm on the same set of independent and dependent variables.
4. Accuracy of 0.467 using the Regression Tree Analysis.
5. The Linear Regression model has a high accuracy of 78%
6. The MinMax accuracy is relatively high, whereas MAPE is moderate to low, which are good indicators.
7. The model has an F Statistic of 48.4 which is mediumly high

the t value of Pitch\_premium, width\_premium, Price\_relative and quality is positive dicating that these predictors are associated with Price\_Premium. A larger t-value indicates that that it is less likely that the coefficient is not equal to zero purely by chance.

Again, as the p-value for Flight\_Duration is less than 0.05 they are both statistically significant in the multiple linear regression model for Price\_Economy response variable.

The model’s, p-value: < 2.2e-16 is also lower than the statistical significance level of 0.05, this indicates that we can safely reject the null hypothesis that the value for the coefficient is zero (or in other words, the predictor variable has no explanatory relationship with the response variable).

1. From the VIF plot we see that Flight Duration and Price Relative are most important factors in predicitng Price Economy.
2. We get an accuracy of 0.447 by implementing the Regression tree model for the Price\_Premium