IST 687 Group 5

Milestone 1

11/06/2019

What we have done

1. Import data to RStudio
2. Using summary() and str() to get the summary information about the data
3. Finding out NAs

By using sum(is.na(df$columnName)) to find what column contains the NA values

The following column has NA values:

* 1. Departure.Delay.in.Minutes (192)
  2. Arrival.Delay.in.Minutes (220)
  3. Flight.time.in.minutes (220)
  4. freeText (1000)

1. Replacing NA values in each column
2. With Flight.time.in.minutes, we ran a linear regression between flight time and flight distance. Since there is no NA value in column Flight.Distance, we replaced the missing values with predictions from our linear regression m0
3. o？del.
4. We will leave freeText as it is since it is additional comments.

Our next step

1. Dealing with categorical variables (i.e., gender, class, airline status, flight cancelled, etc,.)
2. Figure out how to replace NA value in Departure.Delay.in.Minutes and Arrival.Delay.in.Minutes
3. Use correlation method to find which parts of independent variables are more important than others for our target feature (satisfaction)
4. Analyze important independent variables one by one by visualization (plot )and statistic methods (mean,variance, distribution)
5. Split our customers by three levels of NPS (Likelihood to recommend) by score 0-6,7-8 and 9-10. We can analyze different groups to see why people in 0-7 tend to give lower rating; and why people in 9-10 tend to give higher ratings.
6. Split our customers by their flight classes as customers in different classes tend to have different level of expectation for the services. We can study what customers value the most in different classes.

Questions

1. How to replace NA value in Departure.Delay.in.Minutes and Arrival.Delay.in.Minutes.

How is the team working

Everyone is great!