Aaron Chan

STAT 425

5/9/2019

STAT 425 Final Project

The objective of this data analysis is to be able to predict a certain outcome given a set of data. In this case, we are using hotel booking data given to us by our professor but taken from the paper: “Hotel Booking Demand Datasets”. This data contains many categories of interest pertaining to hotel bookings, such as: average daily rate (*adr*), number of nights stayed, number of children/adults, and arrival times, just to name a few. There are over 400 rows of data for out specific dataset pertaining to “Resort Hotel”, we should have enough data to accurately predict a certain variable if we choose to do so. In my case, I would like use average daily rate (*adr*) as a response variable since predicting future earnings depending on customer traits has some very useful real-world applications.

Right off the bat, we can see that some variables such as *is\_canceled*, *arrival\_date\_month*, *meal*, *market\_segment*, *reserved\_room\_type*, and *customer\_type* are categorical variables. The number of occupants: *adults*, *children*, *babies*, can also be considered categorical variables as they are limited in value by available room sizes. Numeric variables in our data set are the remaining variables. Keep in mind that this data only pertains to data with *hotel* type “Resort Hotel”. In this analysis, I decided not to keep variables such as *arrival\_date\_year*, *arrival\_date\_month*,and *arrival\_date\_week*, because *arrival\_date\_month* should be a good enough indication for seasonality. Assuming that economic conditions are similar each year, the month of the stay at “Resort Hotel” should be a good enough indicator. Introduction of other unnecessary time variables may increase autocorrelation between variables. Interaction terms are needed between children and adults as well as babies and adults, for obvious reasons (children and babies cannot book hotel rooms by themselves). After graphing some of the numerical data against *adr* in Figure 1, we do see some evidence against linear trends, nothing in these graphs suggest a strictly linear relationship between explanatories and the response.

INSERT FIGURE 1

We will be using *adr* as our response variable, and all other variables as explanatory.