Analysis of Resort Hotel

Minji Song

Introduction

In this project, I used Hotel Booking Demand data which is from the course website. This dataset includes 2097 American visitor's hotel booking information. The dataset includes 18 variables. There are 6 categorical variables describing booking information of guests such as hotel, arrive date month, meal, market segment, reserved room type, and customer type. Meal indicates the type of meal booked. BB is Bed & breakfast, HB is breakfast and one other meal, FB is breakfast, lunch, and dinner and SC/Undefined indicates no meal package. Market segment has two types: Online TA, TO, direct, corporate, and groups. TA means Travel Agents and TO means Tour Operator. Customer type has four different types: Contract, group, transient, transient party. There are also 13 categorical variables including the value indicating if the booking was canceled, the number of days between booking date and arrival date, year of the arrival date, day of the arrival date, week number of the arrival date, the number of stays in weekend nights, the number of satay in weeknights, number of adults, children, and babies, average daily rate, a total of special requests.

Exploratory Data Analysis

Since I am in session 1, I analyzed corresponding to Resort Hotel which contains 479 guests' information. Thus, I split the data only for Resort Hotel through Excel and then split the data into training data and test data. My goal of this project is to determine if there is any relationship between the response variable and other predictor variables. I chose adr, average daily rate, variable as a response variable because it is important for a hotel to analyze which factors affect the most on a daily rate which is directly related to the sales. Before I do anything first, I changed some values of the is_canceled variable and meal variable. is_canceled variable is a categorical variable with 0 and 1, so I changed it to 0 to No and 1 to Yes to identify values easily. For meal variables, SC and Undefined have the same meaning but they are different values, thus I combined both values to name it as no meal. After I changed some values of categorical variables, I performed Explanatory Data Analysis by plotting each categorical variable to see the relationship with adr. As we look at Figure 1, five bar plots of categorical variables show the proportion of each variable. Meal and reserved room types are right-skewed and the market segment is left-skewed. There is no variable that has a perfect normal graph. In Figure 2, I plotted a box plot for categorical variables as well. For the arrival date month variable, we can see that August is associated with high adr, and January has low adr. Reserved room type H has a higher adr with about 180 adr. Other variables in this plot have similar trend. If we look at Figure 3, these plots show the relationship between continuous variables and adr. It seems that lead time and arrival date week numbers have distinct trend and we can assume that

they can be related to each other somehow and this is the part we need to check by using interaction term. Figure 4 shows correlation plots where we see that arrival date week number and arrival date year tend to have a negative correlation more than other variables. Residual plots indicate if any numerical variables have a nonlinear trend and we can check this in Figure 5. All plots are well distributed and there is no clear trend. Even though lead time plot looks like to be a heteroscedasticity a bit in a little plot, it is distributed well enough. From the residual plots, none of the numerical variables support the nonlinear trend.

Method

Based on explanatory data analysis, I performed linear regression using all variables and interaction terms with lead time and arrival date week numbers. The model includes 16 variables and one interaction term that I wanted to check if it is significant variable in the model. In a summary of the regression, the p-value of the interaction term is lower than 0.05, so we can conclude this is a significant variable and keep this in the model. Since we have a lot of variables, I performed the AIC model selection to reduce some variables based on the small AIC value. Through the selection, the final model has 14 variables: is canceled, lead time, arrival date year, arrival date month, arrival date week number, stay weeknights, adults, children, meal, market segment, reserved room type, customer type, a total of special requests and the interaction terms. Before we confirm the final model, I performed the VIF test to see if there is any multicollinearity with this model. It seems like there is multicollinearity in this model because some values have high VIF values. The highest VIF is arrival date week number with 81.3846, there are few more variables that have over 10 VIFs. However, those variables that have high VIF values can be ignored since those are either categorical variables with more than two levels or part of interaction terms. Other than those variables, all other variables look good. Thus, this would be my final model. With this final model and variables, I predicted how each predictor influences the response variable by using minimum, median, and maximum number of the predictors in data.

Through the prediction, it has the result that a large number of arrival date week number has a higher average daily rate than a small number of arrival date week number. When the arrival date week number is 1, 29 and 53, the predicted adr are 21.74, 103.47, and 173.52. It means the second half of the year has a higher rate. The lead time has the opposite results. The shorter lead time has a higher average daily rate by 30% than 537 days longer lead time. The shorter time between booking date and arrival date have a higher rate. This data set is for between 2015 and 2017. In 2015, the average daily rate is much lower than in 2017. In 2017, the rate has increased by about 45% compared to 2015 and it was -6.5 in 2015. The more nights staying in a week increases the average daily rate. The difference between 0 nights to 10 nights is 28.1%. More adults and children also increase the average daily rate. Also, a lot of special requests increase the daily average rate. A special request increases the rate by about 4.5%.

The next regression I performed is regression trees using rpart function. This regression split the model into several steps and the first node of the variable we split on is arrival date month with 181 observations. The second branch is also an arrival date month with 202 observations. It is

telling that the most important variable in this regression is the arrival date month. This tree plot can be seen in Figure 6. As we look at the size of the tree plot with 11 terminal nodes in Figure 7, error reduction diminish as the tree grows deeper. We can fit a bigger tree by pruning the tree using cp=0.001 instead of cp=0.01. To determine the optimal tree size, we can pick either optimal tree size with the smallest Cross-Validation error or smallest tree size with CV error within 1 standard error of the smallest CV error. In this model, I chose the second option and the red line of the plot in Figure 8 indicates the CV error within 1 standard error of the smallest CV error. Figure 9 is the smallest tree plot based on the CV error that looks simpler than the first tree plot. The optimal cp value has to be below red line in plot and the final plot for the model can be seen in Figure 10. Using the same predictors as linear regression in a tree model, I fitted the random forests to see how training error and cross-validation test errors are different. The training error is 101966.9 and the test error is 381974.4. Thus, we can see that training error is much smaller than the test error, and the training dataset is more accurate.

Result

This hotel booking demand dataset was analyzed through linear regression analysis and regression tree analysis to interpret the relationship between the average daily rate and other predictors. In linear regression analysis, 14 continuous and categorical variables are selected by AIC model selection and this became the final model. To be more specific, it turns out that a large number of arrival date week increase the average daily rate, which means the second half of the year have higher average daily rate than the first half of the year. For the lead time, the number of days between the date of customer book and the arrival date, shorter lead time have a higher average daily rate than longer lead time. The spontaneous customer would pay a higher rate than a planned customer. The more nights customers stay in a week, the higher the average daily rate they have to pay. Also, more adults or more babies have to pay a higher average daily rate. Increasing the total of special requests increase the average daily rate as well.

Through regression tree analysis with the same predictors in a previous regression model, the most important variable in determining the average daily rate is the arrival date month. The customer who visit the hotel in August, July and June tend to pay a higher rate than other periods because those three months are the time people usually go on a vacation. Reserved room type is the second important variable in determining the average daily rate. If the room type is not A or D, it has a higher average daily rate. However, we do not have any details about the room type, so we assume that relatively cheap rooms which is types A and D might have less room quality than other room types. Two analyses I performed have different results, however, this would give the hotel ideas what factors play an important role in determining an average daily rate that is related to increase the sales.

Appendix

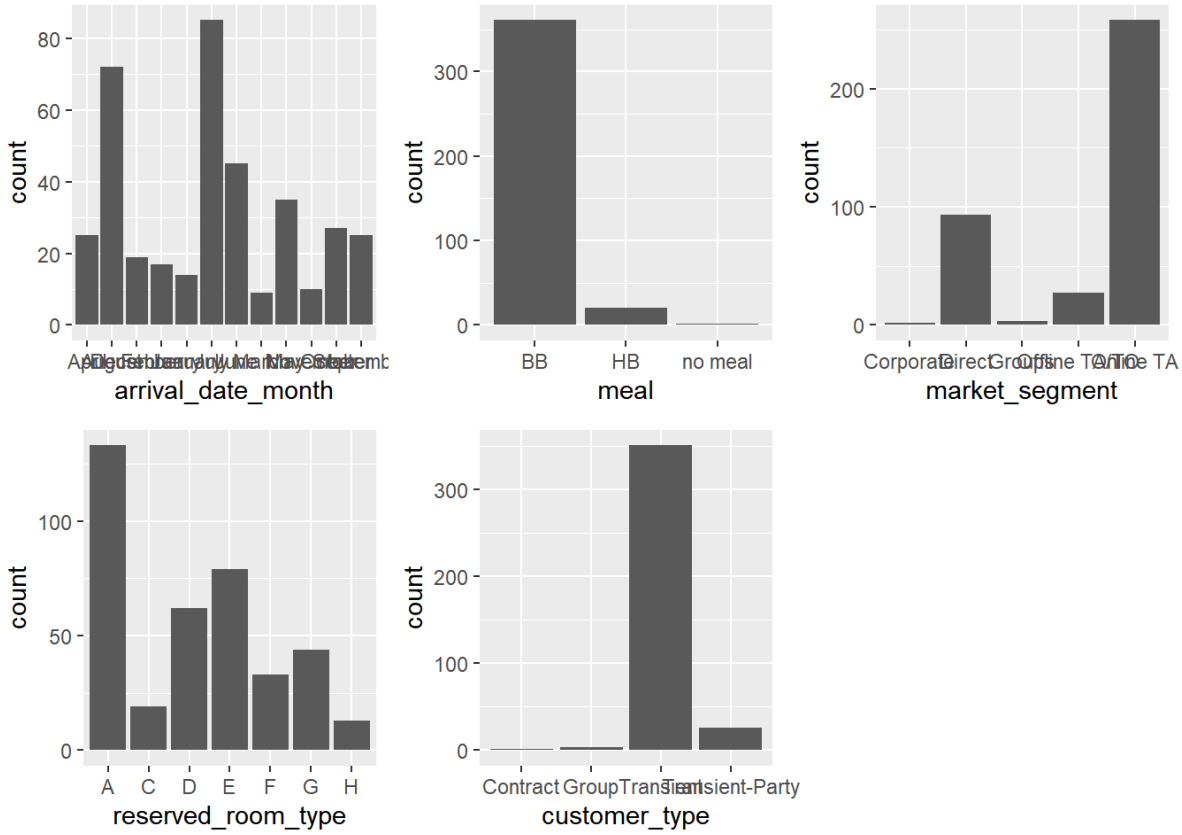

Figure 1

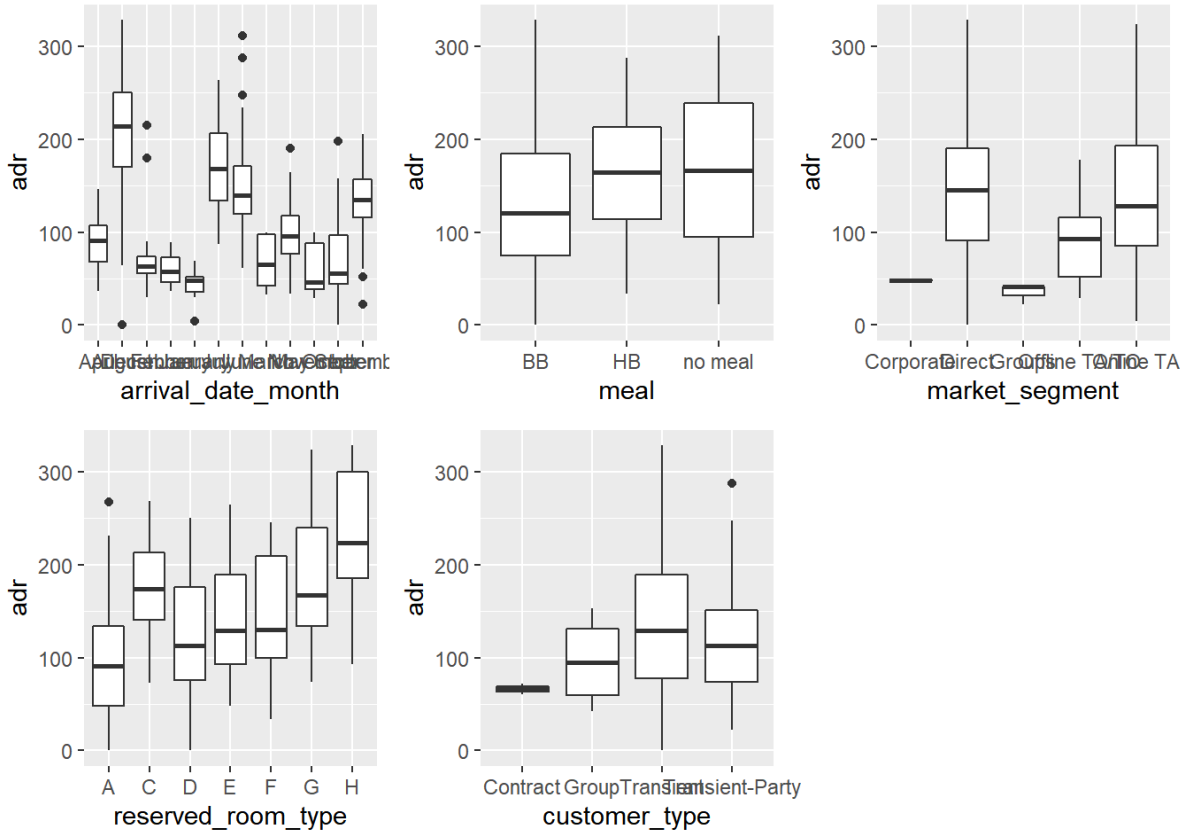
adr
arrival_date_month
meal
BB
HB
no meal
market_segment
Corporate
Direct
Groups
Offline TA/TO
Online TA
reserved_room_type
A
C
D
E
F
G
H
customer_type
Contract
Group
Transient
Transient-Party

Figure 2

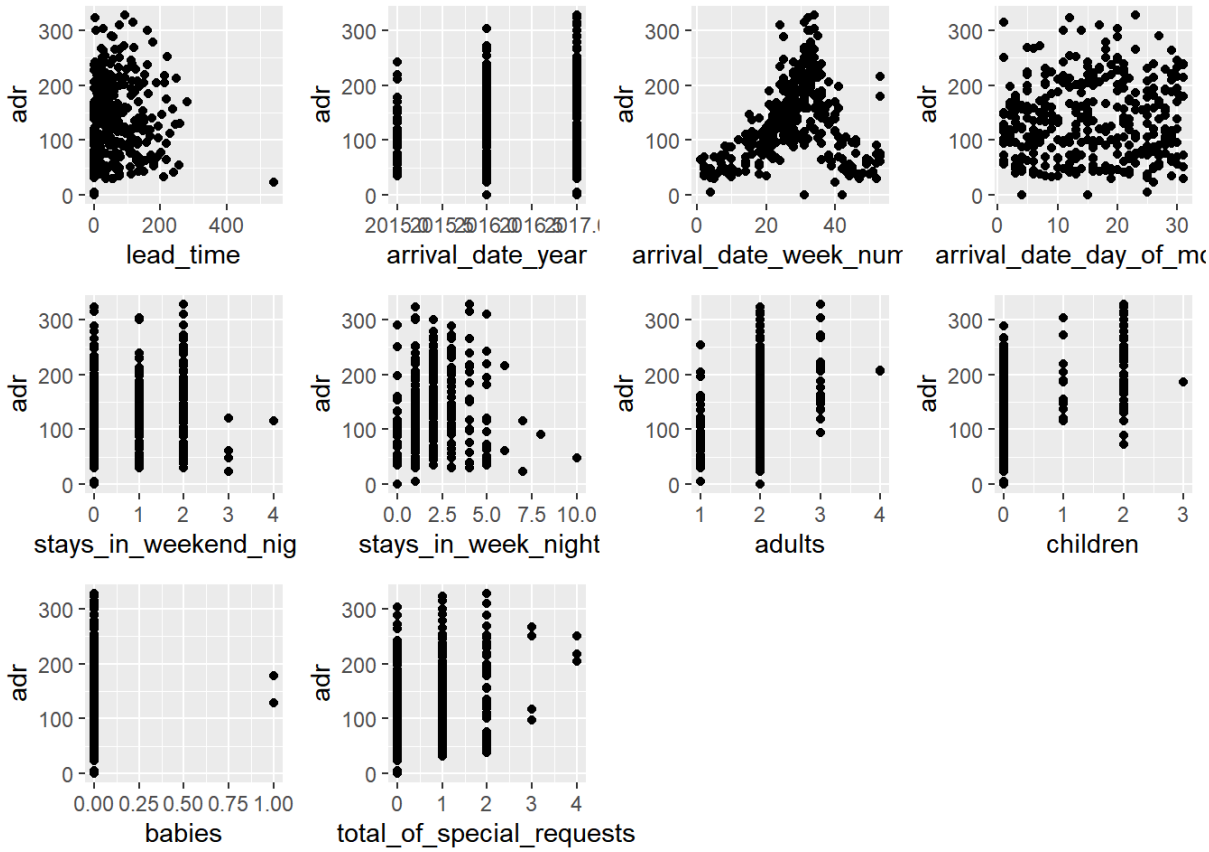
adr
lead_time
arrival_date_year
arrival_date_week_num
arrival_date_day_of_mo
stays_in_weekend_nig
stays_in_week_night
adults
children
babies
total_of_special_requests

Figure 3

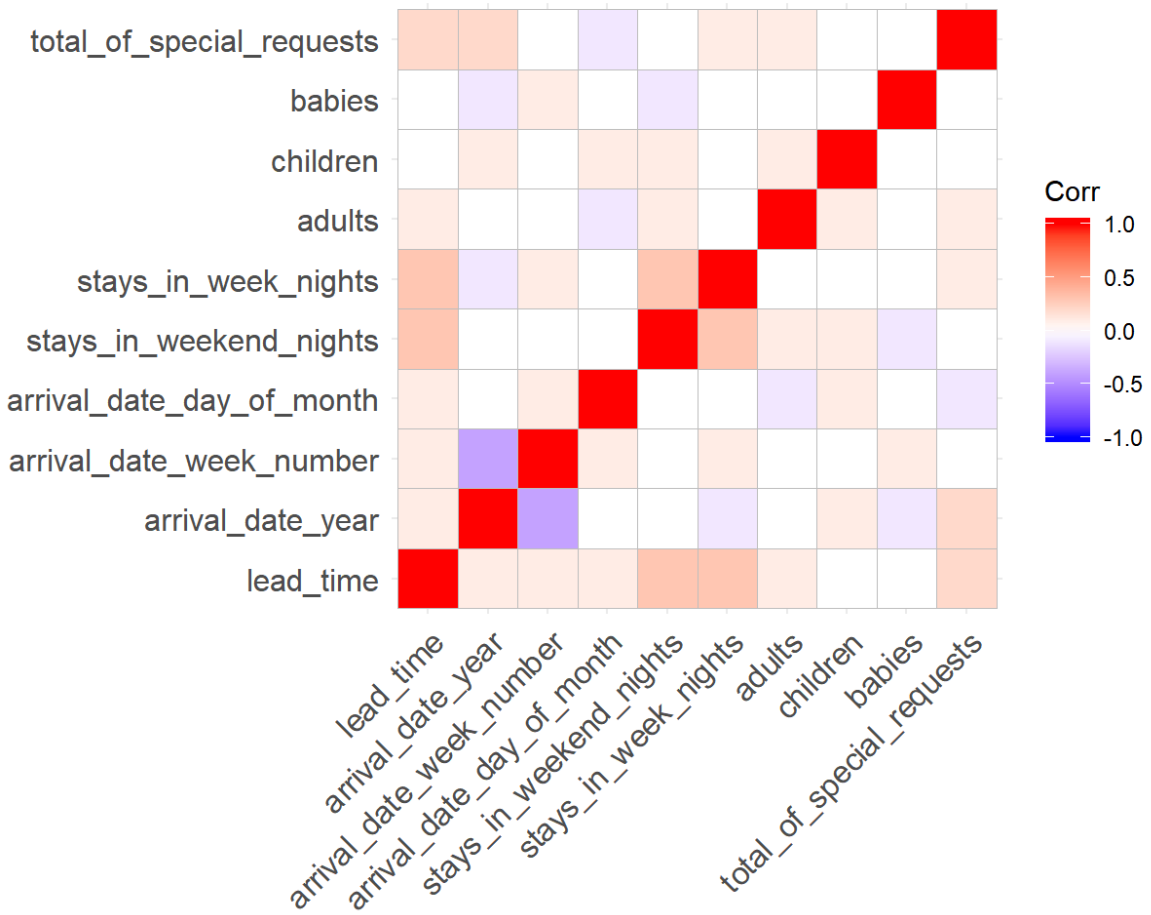
total_of_special_requests
babies
children
adults
stays_in_week_nights
stays_in_weekend_nights
arrival_date_day_of_month
arrival_date_week_number
arrival_date_year
lead_time
lead_time
arrival_date_year
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
total_of_special_requests
Corr
1.0
0.5
0.0
-0.5
-1.0

Figure 4

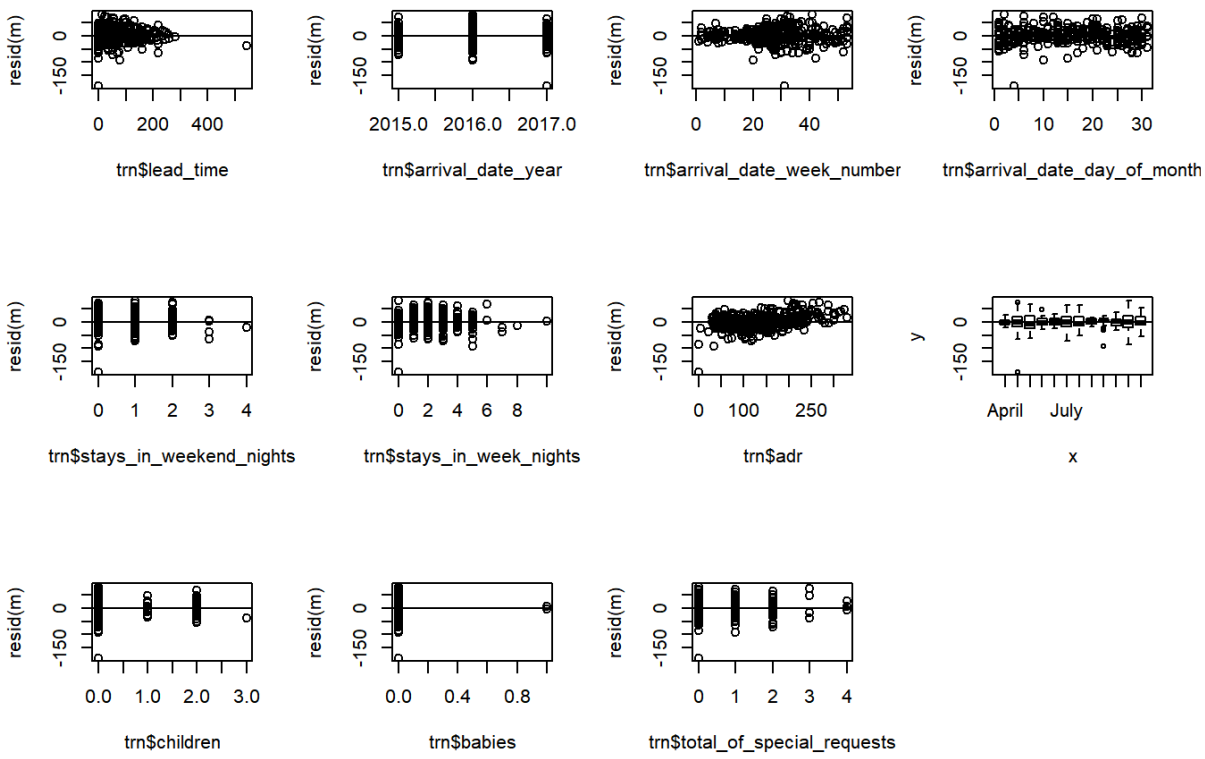

Figure 5

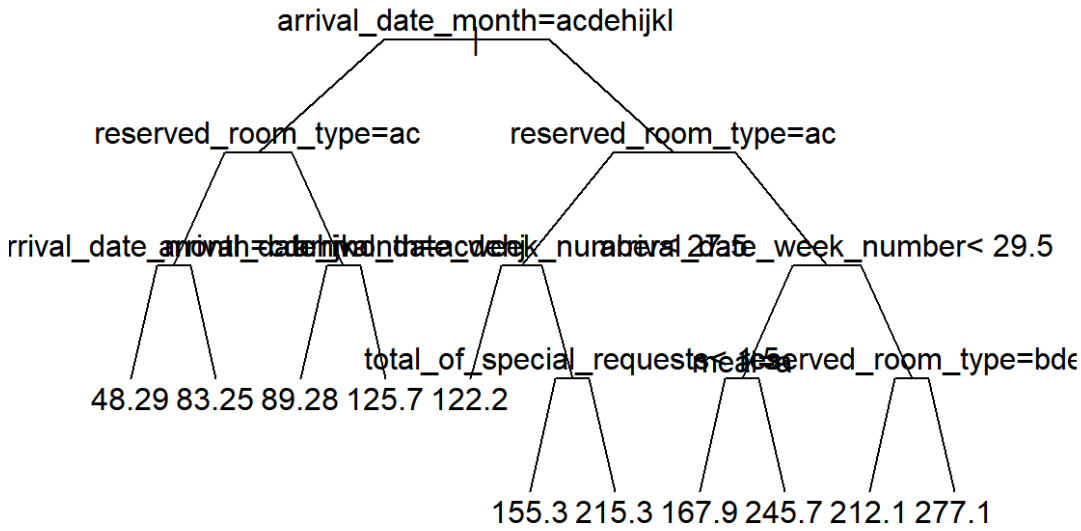

Figure 6

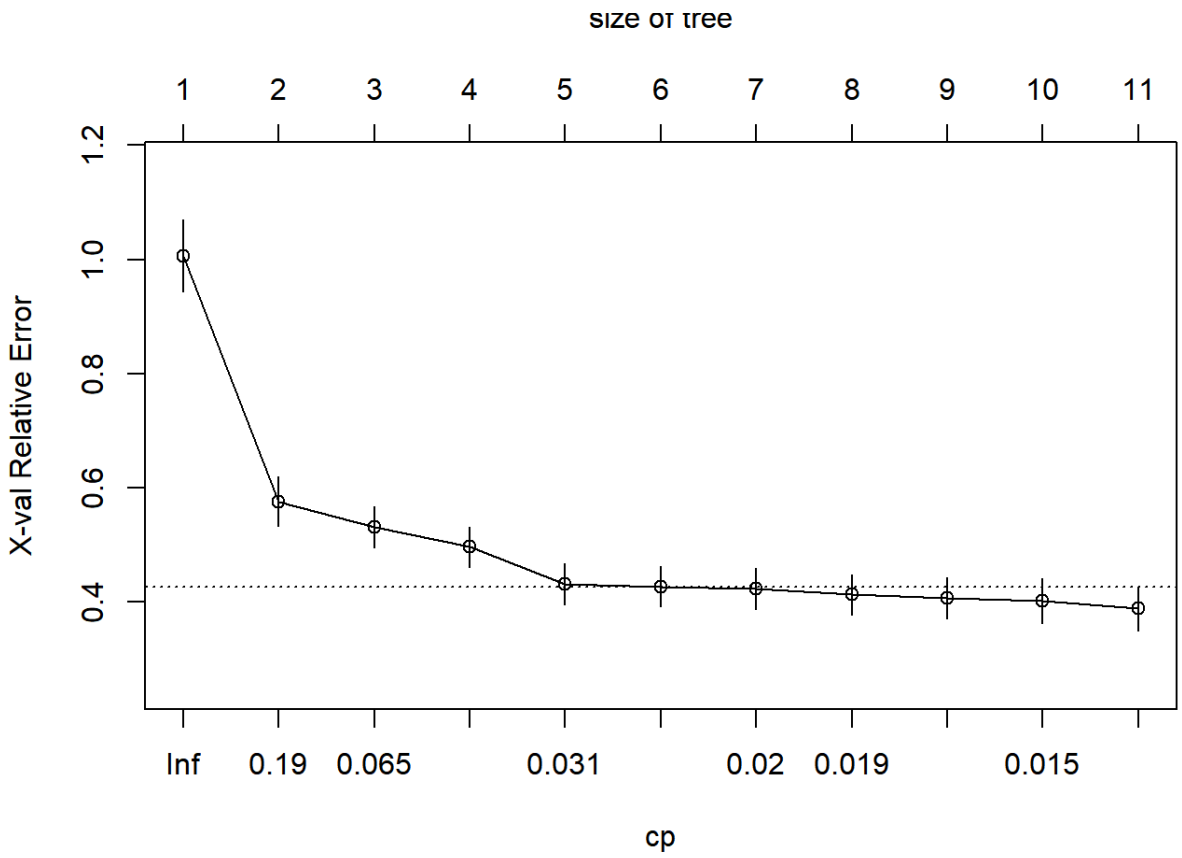
Size of tree
1
2
3
4
5
6
7
8
9
10
11
X-val Relative Error
0.4
0.6
0.8
1.0
1.2
Inf
0.19
0.065
0.031
0.02
0.019
0.015
cp

Figure 7

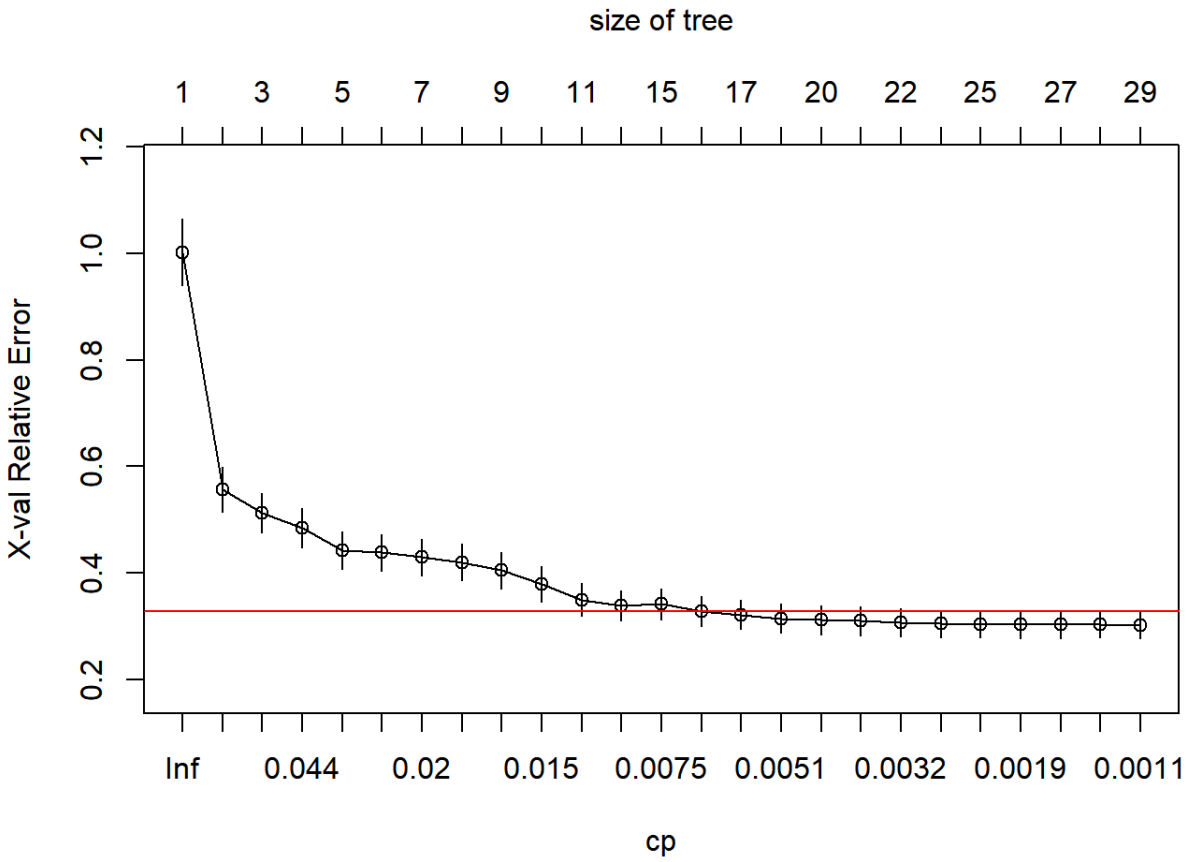
size of tree
1
3
5
7
9
11
15
17
20
22
25
27
29
X-val Relative Error
0.2
0.4
0.6
0.8
1.0
1.2
Inf
0.044
0.02
0.015
0.0075
0.0051
0.0032
0.0019
0.0011
cp

Figure 8

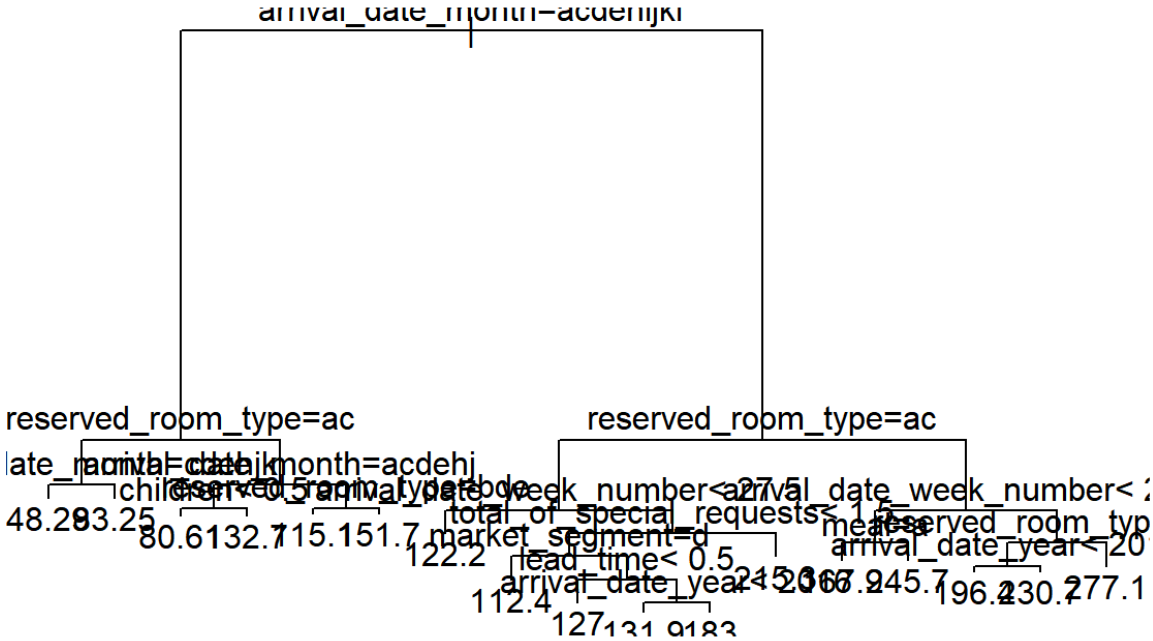
arrival_date_month=acdehijkl
reserved_room_type=ac
reserved_room_type=ac
48.28
93.25
80.61
132.7
115.1
51.7
122.2
112.4
127
131
183
196.2
230.7
277.1

Figure 9

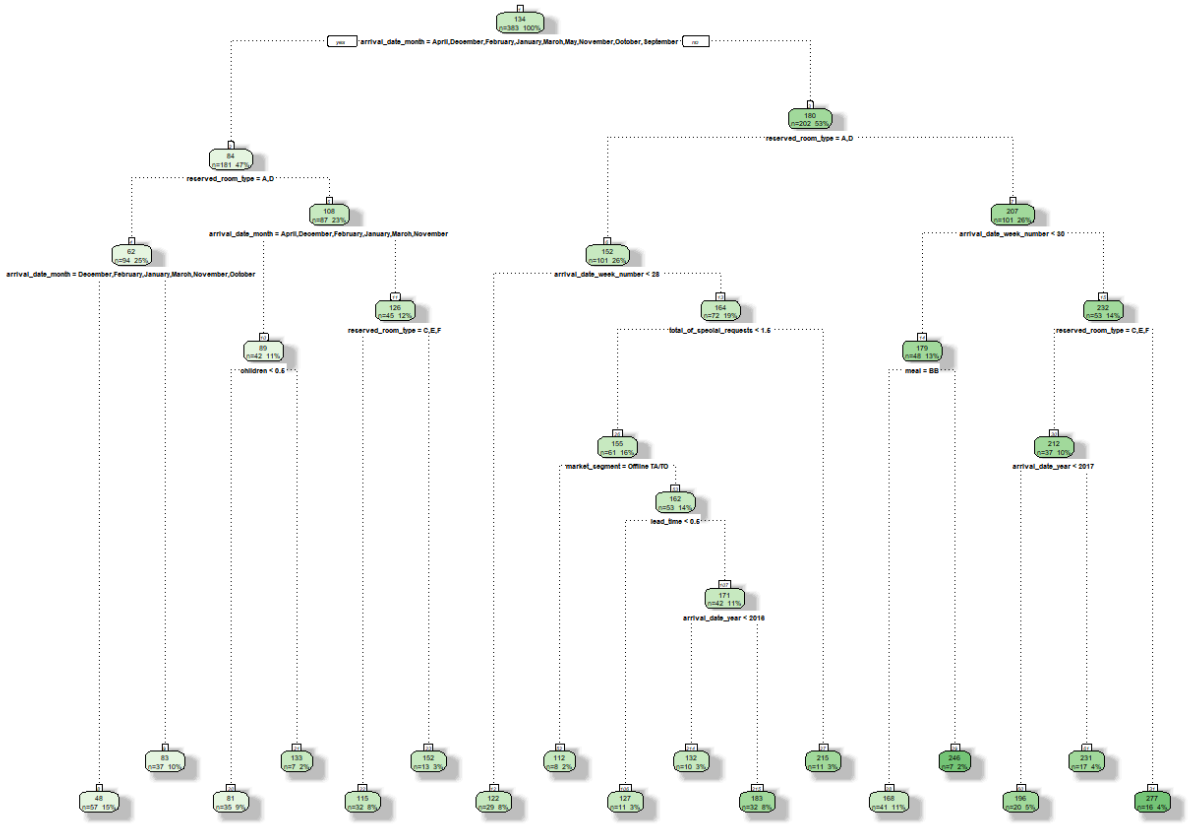
134
n=383 100%
yes
arrival_date_month = April,December,February,January,March,May,November,October,September
no
180
n=202 53%
reserved_room_type = A,D
84
n=181 47%
reserved_room_type = A,D
108
n=87 23%
arrival_date_month = April,December,February,January,March,November
207
n=101 26%
arrival_date_week_number < 30
62
n=94 25%
arrival_date_month = December,February,January,March,November,October
126
n=45 12%
reserved_room_type = C,E,F
89
n=42 11%
children < 0.6
152
n=101 26%
arrival_date_week_number < 28
164
n=72 19%
total_of_special_requests < 1.5
179
n=48 13%
meal = BB
232
n=53 14%
reserved_room_type = C,E,F
155
n=61 16%
market_segment = Offline TA/TO
212
n=37 10%
arrival_date_year < 2017
162
n=53 14%
lead_time < 0.5
171
n=42 11%
arrival_date_year < 2016
46
n=57 15%
83
n=37 10%
91
n=35 9%
139
n=7 2%
115
n=32 8%
152
n=13 3%
122
n=29 8%
112
n=8 2%
127
n=11 3%
192
n=10 3%
183
n=32 8%
215
n=11 3%
168
n=41 11%
244
n=7 2%
196
n=20 5%
231
n=17 4%
277
n=16 4%

Figure 10