STAT425

Final Project Report

Tianqi Wu

1. Introduction

The project aims to predict the average daily rate (adr) for resort hotels based on booking information. The data set is a modified version from Hotel booking demand in Kaggle.[1] The data set contains information like time and length of the stay, meal type and number of special requests. We will fit two linear regression models and one random forest model to make the predictions.

2. Exploratory Data Analysis

The data set has 1618 booking records with 17 variables and there is no missing value. Our response variable is average daily rate and the median value is 112.7. We find that there are 17 observations with adr <= 12. Since they are clearly outliers, we should remove them. In addition, we have three factor levels with only one observation and we remove them as well. The graphic display is presented in Figure 1. For most categorical variables, there are considerable differences across different levels. For most numeric variables, there are positive and negative trends with average daily rate. Further analysis is provided below.

2.1. Which variables are categorical and which are numerical?

There are 7 categorical variables and 10 numeric variables. Year and week number are treated as categorical variables because of the variations observed among different levels. Also, there could be holiday weeks and they are distinct from others with no linear trend.

- categorical: is_canceled, arrival_date_year, arrival_date_week_number, meal, market_segment, reserved_room_type, customer_type
- numeric: lead_time, arrival_date_month, arrival_date_day_of_month, stays_in_weekend_nights, stays_in_week_nights, adults, children, babies, adr, total_of_special_requests

2.2. Should we keep all time related variables in our analysis?

Since rates are related to time of stay. For example, rates may be higher for holidays. Hence, we want to include time related variables in the analysis. However, some of the time related variables may be dependent due to similar information. For example, we may want to use one of day, week number and month.

[1] https://www.kaggle.com/jessemostipak/hotel-booking-demand

2.3. Should we keep week or month in our analysis?

Since week number contains the information of month, we may only keep one of them in the analysis. Results show that month is not significant if we treat week number as categorical variable. Similarly, even if we convert day of month to day of year, chi-squared test concludes that it is dependent of week number. Considering week number would be the best to capture the trend of adr, we only keep it in the analysis.

2.4. For categorical variables, should we include any interactions?

From the interaction plots in Figure 3, there are similar trends for the change of average daily rate across different levels of categorical variables. It indicates that we might need to include the interaction terms in the models.

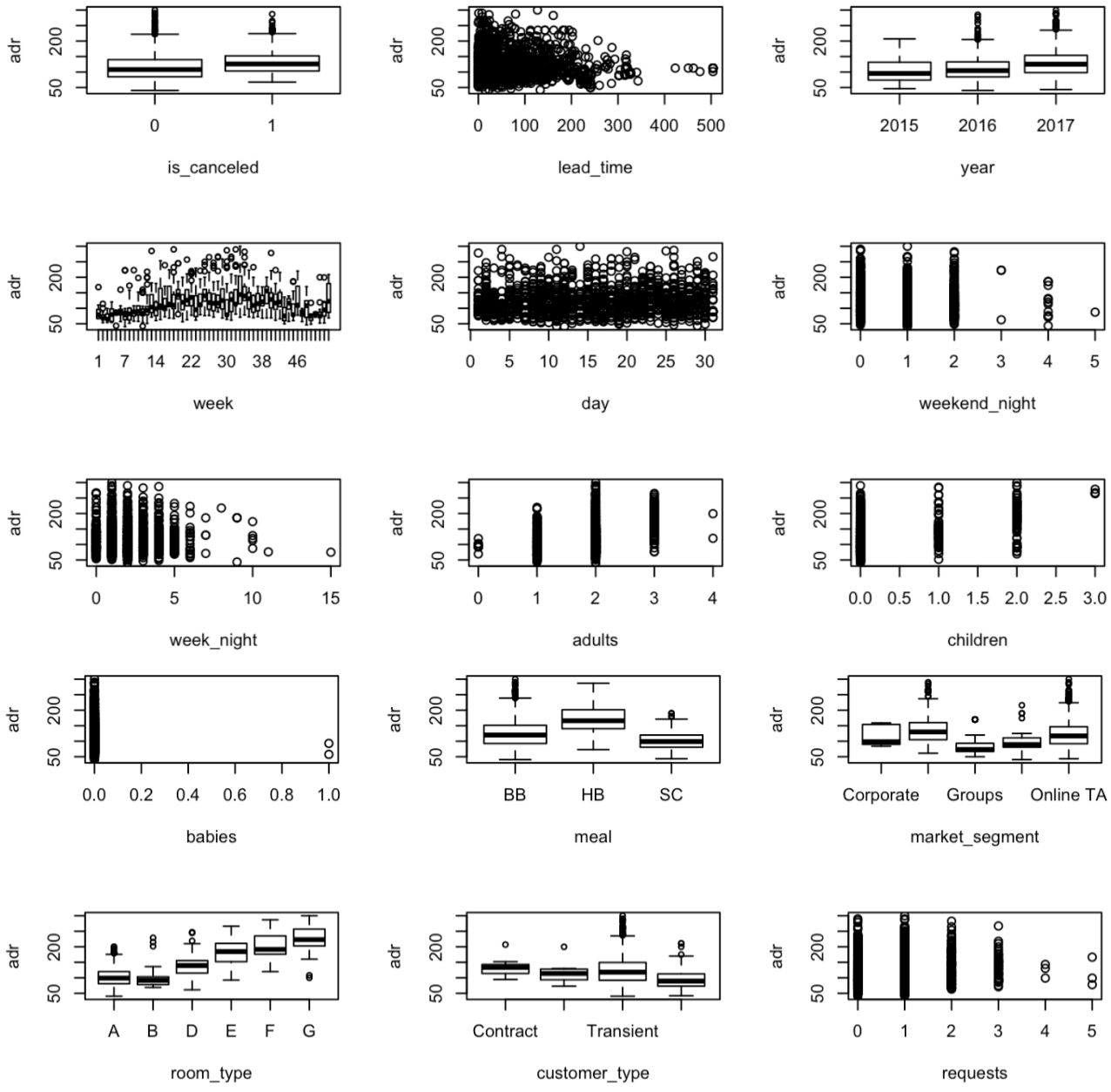

Figure 1: Graphic Display

2.5. For numerical variables, any evidence supporting nonlinear trends?

There are positive trends for number of adults, children. There are negative trends for lead_time, weekend_night, week_night, and number of special requests. Since following analysis shows that quadratic terms of lead_time, number of children and adults are significant, there is evidence of nonlinear trend for those variables.

For categorical variables, there are some levels with very few observations and we may investigate them further. For numerical variables, the correlation plot is provided in Figure 2. Since there is no high correlation, collinearity is not a problem in this case.

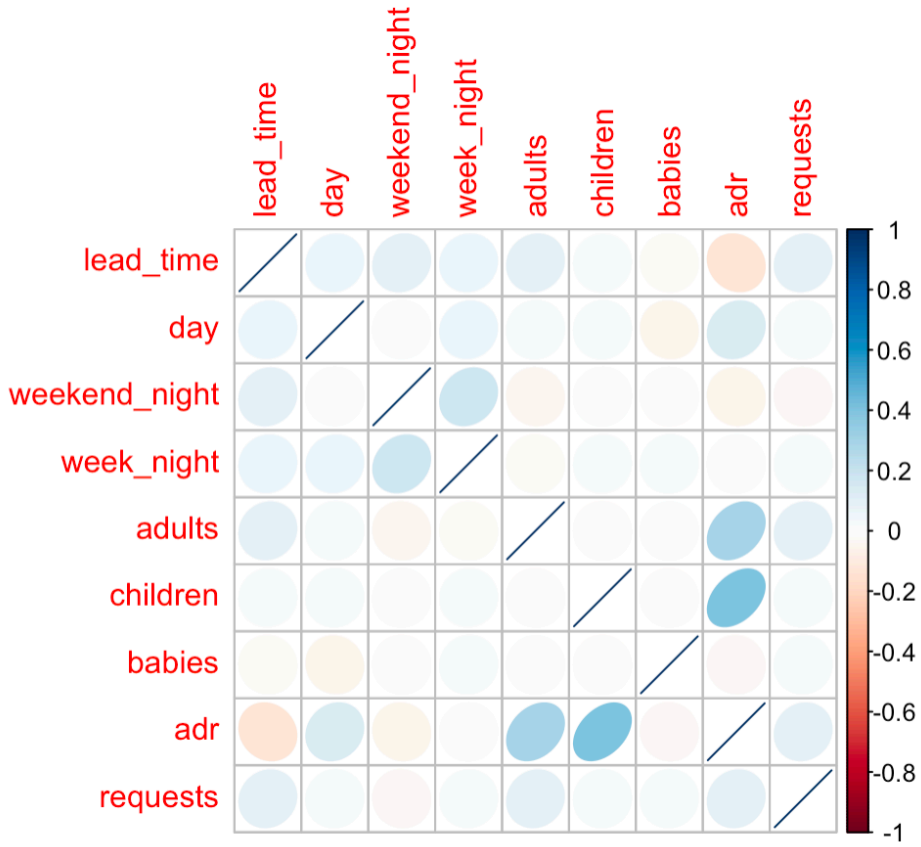

Figure 2: Correlation Plot

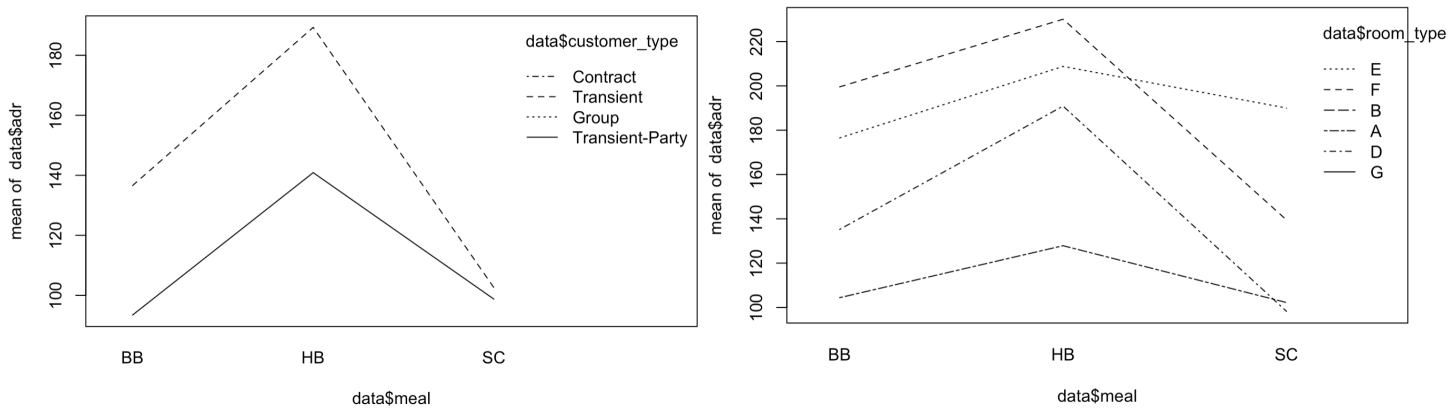

Figure 3: Interaction Plot

3. Method

We compare three models for the prediction task and they are illustrated following. We also split the data into 80% for training and 20% for testing. For evaluation metric of prediction error, we use R-squared and RSME. Equation is given below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \qquad RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$

3.1. Simple Model

We first start with a linear regression model with all predictors and it is illustrated in the equation below. It assumes that mean function $E(y_i)$ is linear in the p predictors and errors e_i are uncorrelated with mean 0 and constant variance. From the model results, the training R-squared is 0.7701 and testing R-squared is 0.725059.

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + e_i$$

$$(\beta_1, \cdots, \beta_p, \sigma^2) \quad : \quad \text{the unknown but true parameters.}$$

$$e_i's \quad : \quad \text{random errors.}$$

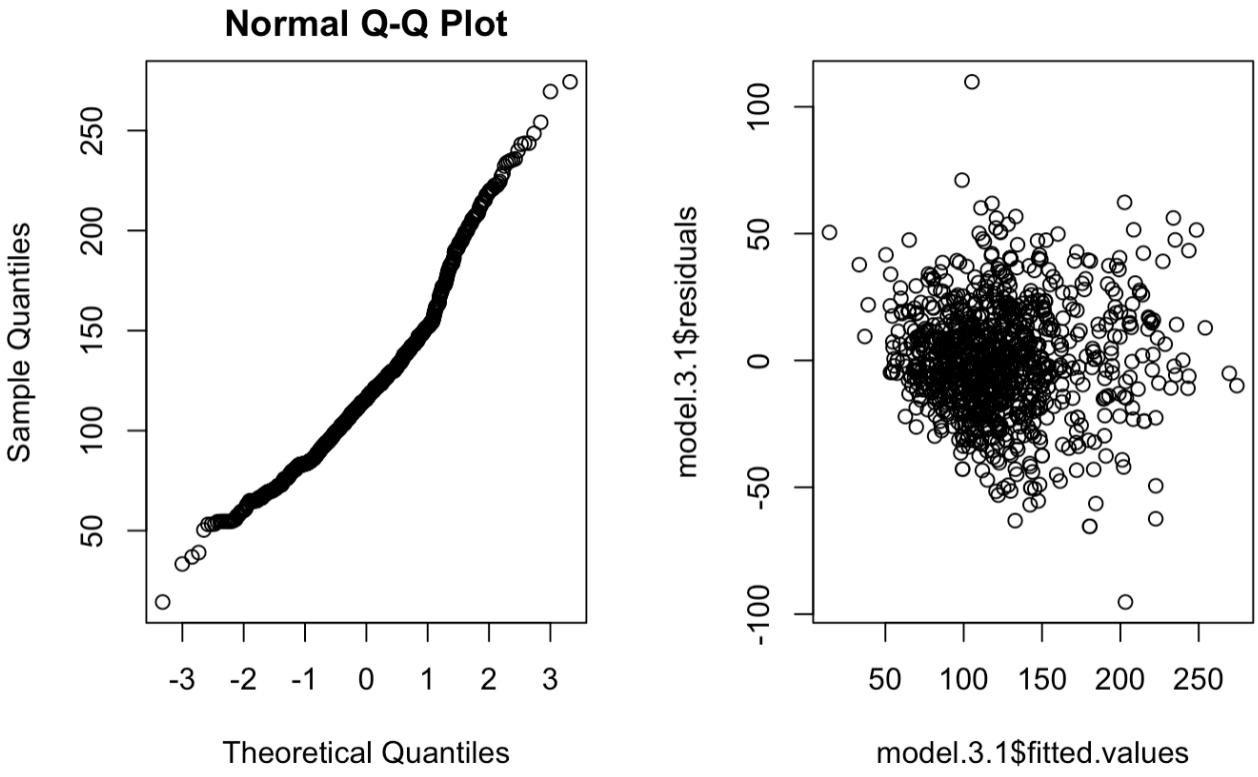

Figure 4: Diagnostics Plots

Then, we perform model diagnostics for the model in 3.1. Breusch-Pagan Test and Shapiro-Wilk test suggest that constant variance and normality assumptions are violated. QQ-plot and residuals vs. fitted plot are presented in Figure 4. The points are not very linear for QQ-plot and the variance is not constant. Durbin-Watson test suggests that errors are independent. The partial residual plots in Figure 5 indicate the linear relationship is not violated.

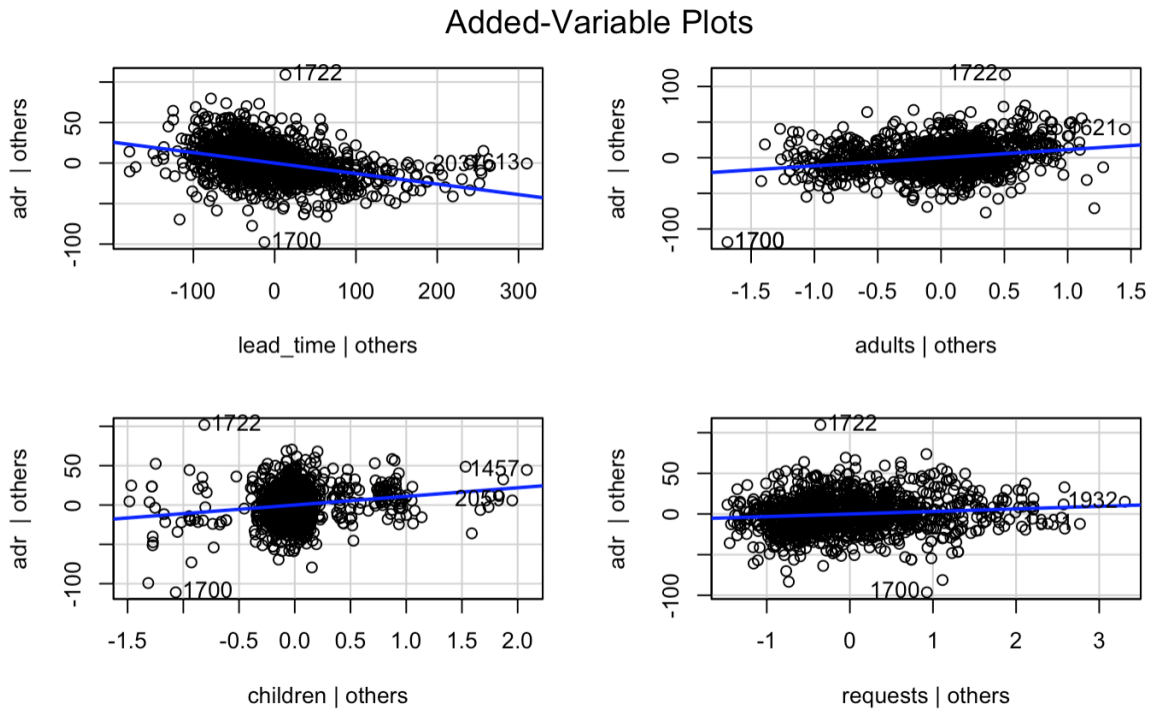

Figure 5: Partial Residual

3.2. Linear Regression Model

For this model, we first use AIC criterion to select significant variables for model in 3.1. The resulting variables are: is_canceled + lead_time + year + week + adults + children + meal + market_segment + room_type + customer_type + requests

Furthermore, we perform tests to include some significant interaction terms and quadratic terms.

- interaction terms: meal:market_segment + meal:room_type + meal:requests + market_segment:room_type
- quadratic terms: I(lead_time^2) + I(children^2) + I(adults^2)

Then, we perform model diagnostics again. There are some unusual observations. '510', ‘1061' and ‘1487' are high leverage point. ‘1722’, ’1700’ and ‘1575’ are outliers. Since they deviate from the data a lot, we remove them. Checking of condition number and variance inflation factor suggests that collinearity is not a problem.

After refitting the model with new training data, Box-Cox transformation is used and Figure 6 suggests that the optimal λ is around 0.38. Other transformations on predictors like log, inverse and square have been attempted but they are not improving the model.

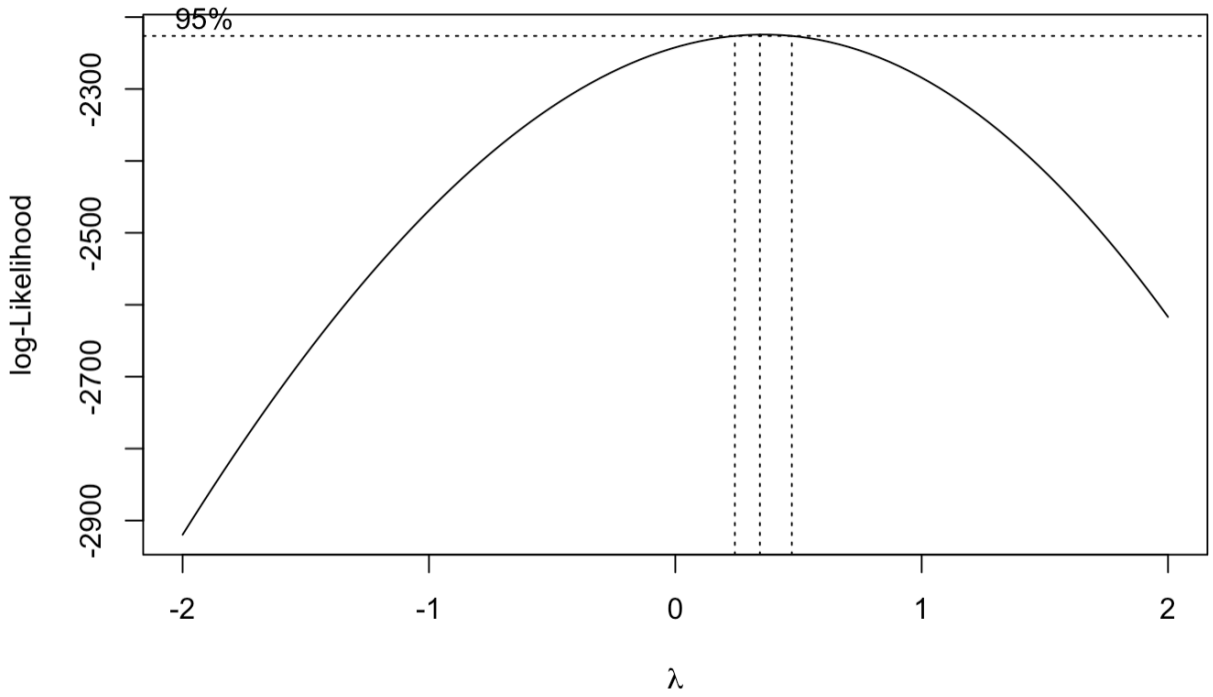

Figure 6: Box-Cox

Finally, we refit the linear regression model and the training R-squared is 0.8073 and testing R-squared is 0.7426. Although Breusch-Pagan Test and Shapiro-Wilk test suggest that constant variance and normality assumptions are violated, QQ-plot and residuals vs. fitted plot are improved in Figure 7.

3.3. Random Forest Model

Random forest model is also used as prediction model and it is a kind of ensemble learning method. First, we grow 500 regression trees and use the average of their predictions as the model prediction. In addition, each tree is built based on a random subset of the data and each split is selected on a random subset of variables. Default parameters of randomforest( ) are used from randomForest library (version 4.6-14). The training R-squared for this model is 0.7550 and testing R-squared is 0.7245.

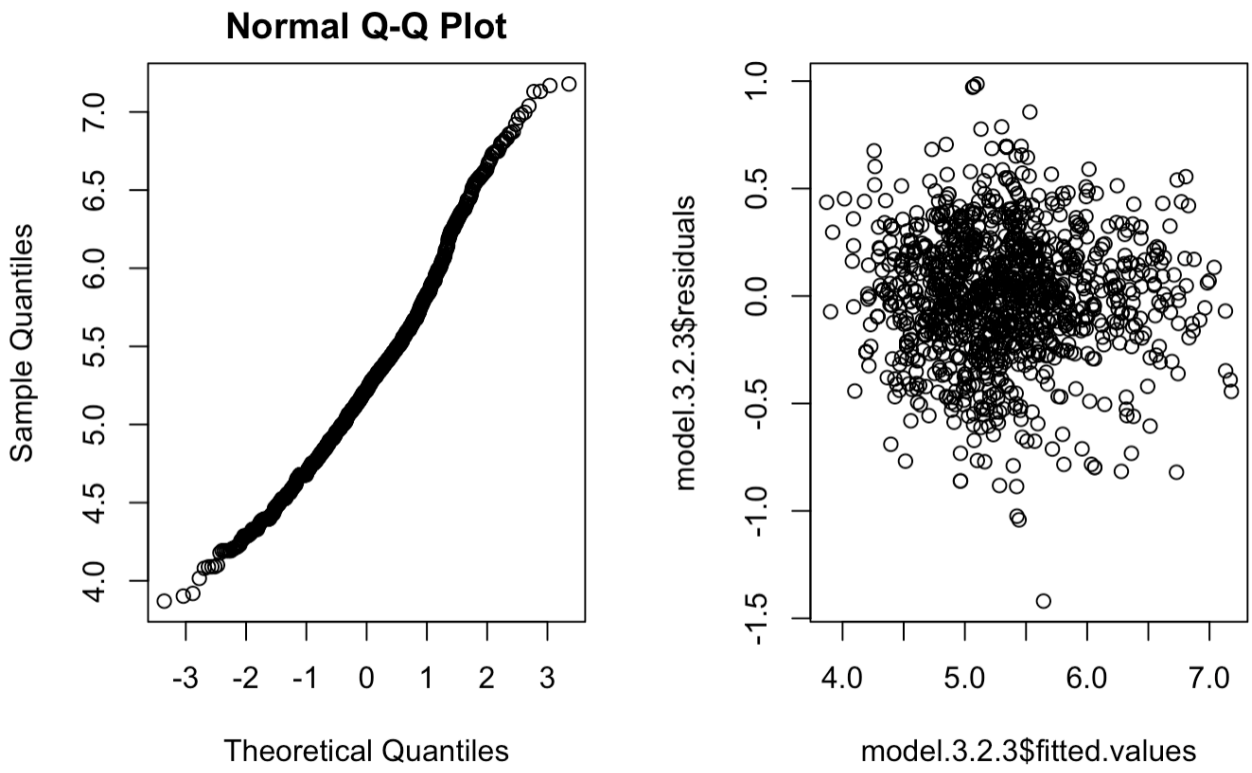

Figure 7: Diagnostic Plots

4. Discussion of Results

The results are shown in Table 1. Model performance from worst to best is random forest < simple < linear regression. Since default random forest is used, the performance could be improved with more parameter tuning. Box-Cox transformation, variable selection, removing outliers, adding interaction and quadratic terms altogether improve the performance of linear regression model. With the efforts done, linear regression model is finally able to outperform random forest for both R-squared and RMSE. If the linear trend of data is not clear, random forest model is likely to get better performance.

Method	Training R^2	Testing R^2	Training RMSE	Testing RMSE
3.1 Simple	0.7701	0.7250	20.907	23.3545
3.2 Linear Regression	**0.8073**	**0.7426**	**18.4951**	**22.5933**
3.3 Random Froest	0.7550	0.7245	21.5833	23.3742

The testing R-squared for best model is 0.7426. It means 74.26% of the variation in the testing data is explained by the model and it indicates that our final model fits data well. For numeric variables, lead_time and number of adults are negatively related to average daily rates, indicating decreasing of those variables for increasing rates. Whereas, number of children and special requests are positively related to average daily rates.

For categorical variables, the average daily rates are expected to be higher if year=2017, week > 8, meal=HB, room_type=G, customer_type=Transient-Party. Although some of the interactions terms are not estimated due to lack of data for these combinations, there are several significant interaction terms like mealSC:market_segmentDirect, mealSC:room_typeD and mealSC:requests.

Despite relatively high R-squared for the linear regression model, there is room of improvement. Currently, the normality and constant variance assumptions are violated even with Box-Cox transformation. Better transformation may be needed to reveal the linear trend of data. Also, more complex models may be used to better fit the data.