Airbnb Price Prediction using Multiple Linear Regression

Christian Jefferson Halim

12/6/2022

1.Introduction

1.1 Research Question

What variables contribute to the dynamic pricing in the Airbnb rental price?

1.2 Background

Airbnb has been the leading company for lodging and tourism experiences worldwide. Founded in 2007, Airbnb has grown to have over 6 million listings in more than 100,000 cities worldwide, making it a leading source of alternative lodging options. One of the critical factors in Airbnb's success is its dynamic pricing model, which allows hosts to adjust their rates based on demand and availability. The price is the key factor in ensuring that it can benefit both the tenants and the customers.

1.3 Purpose

The Airbnb Dynamic pricing introduces a way that two houses next to each other can have different prices. This is because each listing is determined by various independent variables. These can include the location, size, property type (such as a house or apartment), the amenities offered by the property, etc. This paper will help to predict whether those factors come into consideration with the listing price.

1.4 Correlation with Previous Paper

This paper has some characteristics similar to other papers mentioned in phase 1. The previous papers used some similar approaches, such as calculating the mean squared error and implementing a multiple regression model, which is also implemented in this paper. However, I used a different dataset and different methods of selecting the appropriate model with Forward/ Backward Stepwise Regression of AIC, BIC, and Lasso/ Group Lasso.

2. Methods

2.1 Data Cleaning, Split Data, and Removing Outlier

From the initial data, I filtered the data by selecting the 'Church-Yonge Corridor' neighborhood as my main focus, and I chose five predictor variables, which are 'room_type,' 'accommodates', 'beds', 'review_scores_rating', 'host_is_superhost'. I created two dummy variables for the categorical variables: 'room_type_dummy' and 'superhost_dummy'. I also split the dataset into two parts, a training dataset and a test dataset, which consists of a random sampling of the dataset with a percentage of 70% and 30& respectively. A training dataset will be used to train the model, while the test dataset will be used to measure the model's accuracy. Using the IQR formula, I successfully removed 24 outliers and decreased the dataset from 238 datasets to 214 datasets.

2.2 Checking Assumption

The first step in building and choosing the best regression model is to check all the linear regression assumptions. By creating a scatterplot for each predictor variable, I want to see a linear relationship between the price variable and each predictor variable. For the following three assumptions, I will create multiple models first, which consist of:

$$y = \beta_0 + \beta_1\ x_{room\_type} + \beta_2\ x_{accommodates} + \beta_3\ x_{beds} + \beta_4\ x_{review\_scores\_rating} + \beta_5\ x_{superhost} + \epsilon$$

I can analyze the pattern shown in the Standardized Residual vs Fitted of the entire model to satisfy the independence error and homoscedasticity assumption. The homoscedasticity assumption is violated if any pattern is shown, especially a fanning pattern where the residuals become gradually spread out in the plot. Furthermore, I will create a Normal Q-Q Plot to satisfy the normality assumption. The most crucial thing in this plot is seeing a linear pattern for each point. If there are points that made the plot not linear, I will apply a box-cox transformation by transforming the price to log(price) and formulating the data with the maximum lambda value.

2.3 Checking Multicollinearity and Influential Points

To check multicollinearity, I can use the VIF (Variance Inflation Factors) to see whether two or more independent variables are highly correlated, which means the VIF score has to be lower than 5 to fulfill the multicollinearity. Furthermore, I need to check the influential points in the training dataset using Cook's Distance, DFFITS, and DFBETAS. If any influential points are presented in the plot, I need to remove the points by using one of either of the three methods specified.

2.4 Model Selection

This model combination will help us to determine the confidence and prediction interval, which will be calculated in the last process. I used different types of Forward/ Backward Stepwise Regression of AIC, BIC, and Group Lasso to determine the best model. Each method will give a combination of predictor variables that provide the mean squared error, mean absolute error, and prediction error. The method that will give the lowest value in each error will be chosen.

2.5 Checking Variable's Assumption after Stepwise Regression

I need to verify the assumption of the updated model with the same approach as 2.2.

2.6 ANCOVA/ANOVA Test

Since I have accounted for the covariates of the data, I would like to analyze if there is a statistically significant difference between the three variables that I have chosen from the Group Lasso method. If categorical variables are chosen from the method stated in the previous part, I will use the ANCOVA method instead of ANOVA. Then, I can interpret the result of the ANCOVA/ANOVA table to see whether it is statistically significant from the test statistics. It follows with a computation of the 95% of confidence and prediction interval. I also can count for the prediction error by using a formula $ abs(realvalue-pred)/real value $, where the real value is the value in the testing dataset and pred is the prediction price.

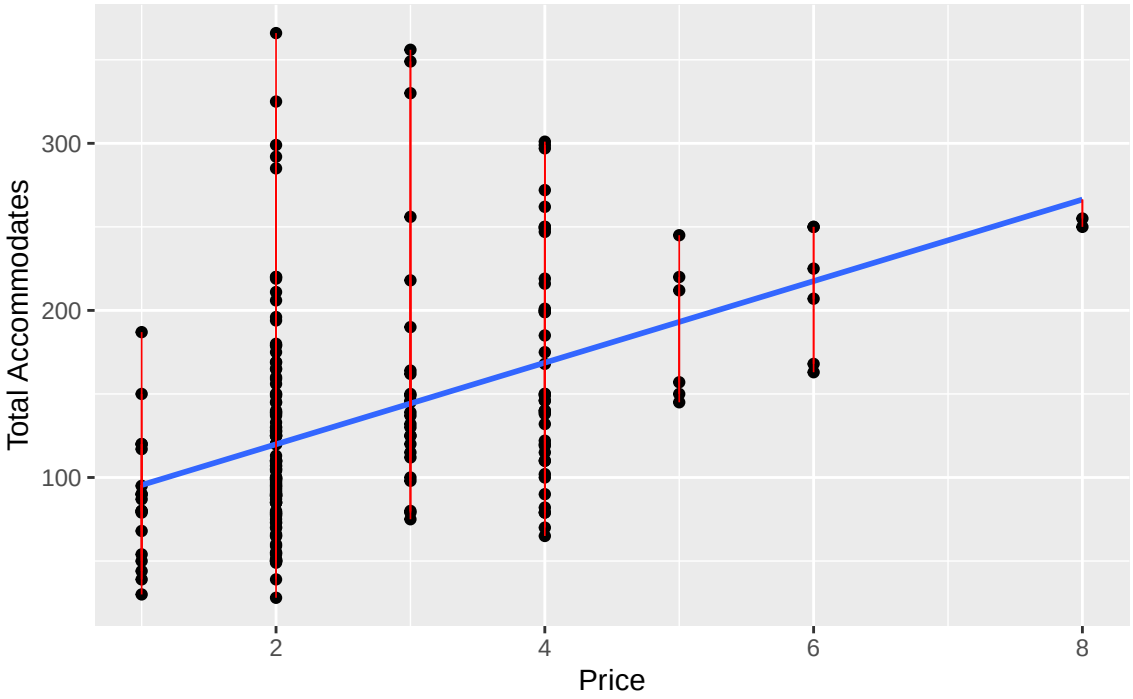
3. Results

3.1 Assumptions

Upon looking at the scatterplot of each predictor variable, all predictor variables fulfilled the linearity assumption since they showed a linear relationship with the price variable. One of the scatterplots of the 'accommodates' variable is shown below. Next, the square root of the standardized residual vs fitted value of the full model shows a pattern, so a box-cox transformation will be used to remove those patterns. After I used the transformation using log(price) and maximum lambda, the graph showed no pattern, and the

residuals had a pretty flat line, such that the independence and homoscedasticity assumptions were met. On the Normal Q-Q plot after the transformation, I can see a linear pattern for all of the points, so the normality assumption has also been met.

Figure 1: Total Accommodates vs Price



3.2 Multicollinearity and Influential Points

Using the VIF method, each predictor variable has a VIF score lower than 5, so I can keep all the variables in our model. Furthermore, in Cook's Distance graph below, the red points represent the influential points that need to be removed. After I subset the data and removed all the influential points, I have a linear Normal Q-Q plot without any influential points being shown here.

Figure 2: Cook's Distance of the Full Model

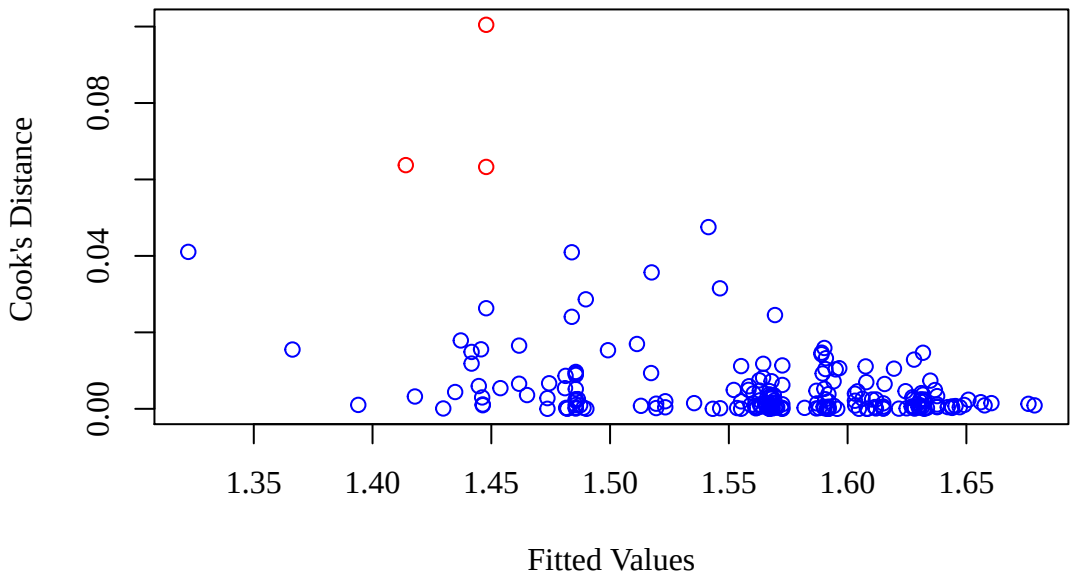
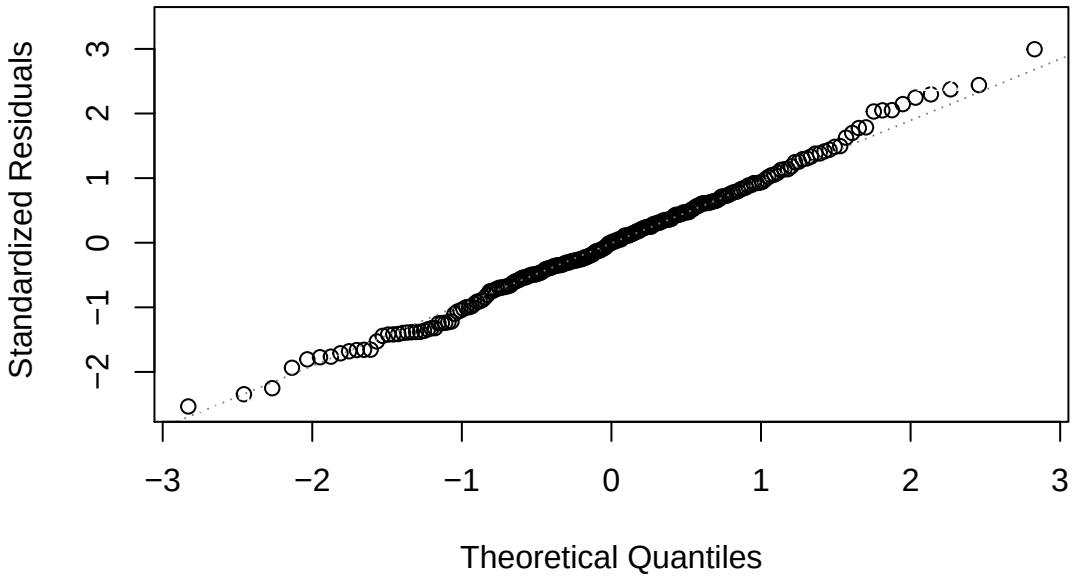


Figure 3: Normal Q−Q Plot without Influential Point



3.3 Model Selection

Upon looking at the prediction error, the mean squared error, and the mean absolute error of each stepwise regression method, the BIC and the Group Lasso method have the same value of the lowest value of the mean

squared error, and the mean absolute error, and the prediction error so I can choose our predictor variables of the method based on either one of the methods. Both the BIC and The Group Lasso recommended using only 3 predictor variables: ' room_type_dummy', 'review_scores_rating', and 'accommodates'.

Table 1: Different Methods of Model Selection

Method	Prediction Error	Mean Absolute Error	Mean Squared Error
AIC	0.01561753	0.09412954	0.01423352
BIC	0.01554620	0.09376275	0.01421103
Group Lasso	0.01554620	0.09376275	0.01421103

3.4 ANCOVA, Confidence Interval, and Prediction Interval

I also need to verify the updated model using a similar approach as in 2.2. Since there is one categorical variable here, I will use the ANCOVA table for our model. Using a p-value of 0.05, the interaction between room_type_dummy and accommodates have a value greater than 0.05, meaning they are insignificant. Thus, I can dispense this interaction and conclude that there is a dependent relationship between room_type_dummy and accommodates. Then, I removed the insignificant variables from the ANCOVA table and created a new one with just the significant variables.

In addition to that, I included a random entry of the training dataset (price = 117, accommodates = 1, review_scores_rating = 4.94). I also get a prediction error of 1.89%, which is considerably small. I also checked with other random entries, and the result is similar.

Table 2: ANCOVA Table with only Significant Variables

	Sum Sq	Df	F value	Pr(>F)
accommodates	0.1915194	1	26.762957	0.0000005
as.factor(room_type_dummy)	0.2044173	1	28.565308	0.0000002
review_scores_rating	0.1098480	1	15.350182	0.0001210
as.factor(room_type_dummy):review_scores_rating	0.0409011	1	5.715522	0.0177028
Residuals	1.4956326	209	NA	NA

4. Discussion

4.1 Influential Points

The influential points can be determined using Cook's Distance, DFFITS, or DFBETAS. Cook's Distance calculates each observation's leverage and residual values, DFFITS calculates the effect of each observation on all of the fitted values, and DFBETAS calculates showed observations that are influential in a given parameter. All methods are calculated in this paper to strengthen the accuracy of the result.

4.2 Model Selection

The Group Lasso method is used in this paper instead of the regular lasso because our model consists of categorical variables. I grouped five observations into a group of 5 and calculated the prediction error based on this group. Since all of the BIC and Group Lasso values are similar, I can choose either method. However, I choose the Group Lasso method for this paper.

4.3 Final Model Analysis

In the final model of the ANCOVA, I have three predictor variables and one dependent variable between review_scores_rating and room_type_dummy. The interpretation of the model is for every increase in the number of accommodates, the price increase by approximately 0.166. Furthermore, the price increased by

0.196 if the room of Airbnb is private. The price increased by 0.066 and 0.03 for every increase of 1.0 and 1.0 in review_scores_rating and as.factor(room_type_dummy):review_scores_rating, respectively.

4.4 Limitations of the Analysis

My model still has a pattern shown in the Residual vs Fitted plot, even after I did a transformation with box-cox transformation. Perhaps another type of transformation can be used here, like a square root or log transformation. The normal Q-Q plot also shows some nonlinearity in the bigger or upper points, which can happen due to some datasets that are not reasonable. All of these violations can affect the overall model performance. Furthermore, my analysis of the stepwise regression with BIC and Lasso showed a similar result, and with such removal of the data points from the argument above, it can also affect the model selection.

5. Bibliography

1. About Us. Airbnb Newsroom. (2022, December 15). Retrieved December 20, 2022, from https://news.airbnb.com/about-us/
2. Inside Airbnb. (n.d.). Retrieved October 20, 2022, from http://insideairbnb.com/get-the-data/
3. Folger, J. (2022, July 12). How airbnb works. Investopedia. Retrieved October 20, 2022, from https://www.investopedia.com/articles/personal-finance/032814/pros-and-cons-using-airbnb.asp#:~:text=Airbnb%20is%20an%20online%20marketplace,some%20income%20from%20their%20property.
4. Librarysearch.library.utoronto.ca. (n.d.). Retrieved October 20, 2022, from https://librarysearch.library.utoronto.ca/discovery/fulldisplay?docid=cdi_elsevier_sciencedirect_doi_10_1016_j_jhtm_2020_08_015&context=PC&vid=01UTORONTO_INST%3AUTORONTO&lang=en&search_scope=UTL_AND_CI&adaptor=Primo+Central&tab=Everything&query=any%2Ccontains%2Cairbnb+price+regression&offset=0
5. Hedonic pricing and the sharing economy: How profile characteristics affect airbnb accommodation prices in Barcelona, Madrid, and Seville. Taylor & Francis. (n.d.). Retrieved October 20, 2022, from https://www.tandfonline.com/doi/full/10.1080/13683500.2020.1718619
6. Customized regression model for Airbnb dynamic pricing. SIGKDD - KDD 2018. (2018, May 18). Retrieved October 20, 2022, from https://www.kdd.org/kdd2018/accepted-papers/view/customized-regression-model-for-airbnb-dynamic-pricing
7. Dye, S. (2020, February 19). Quantile regression. Medium. Retrieved October 20, 2022, from https://towardsdatascience.com/quantile-regression-ff2343c4a03
8. Prabhakaran, S. (n.d.). Assumptions of Linear Regression. 10 Assumptions of Linear Regression - Full List with Examples and Code. Retrieved October 20, 2022, from http://r-statistics.co/Assumptions-of-Linear-Regression.html

6. Appendix

Table 3: VIF Test

	VIF Score
room_type_dummy	1.255146
accommodates	1.939340
beds	1.767281
review_scores_rating	1.038023
superhost_dummy	1.038343

Figure 4: Normal Q−Q Plot with Outliers

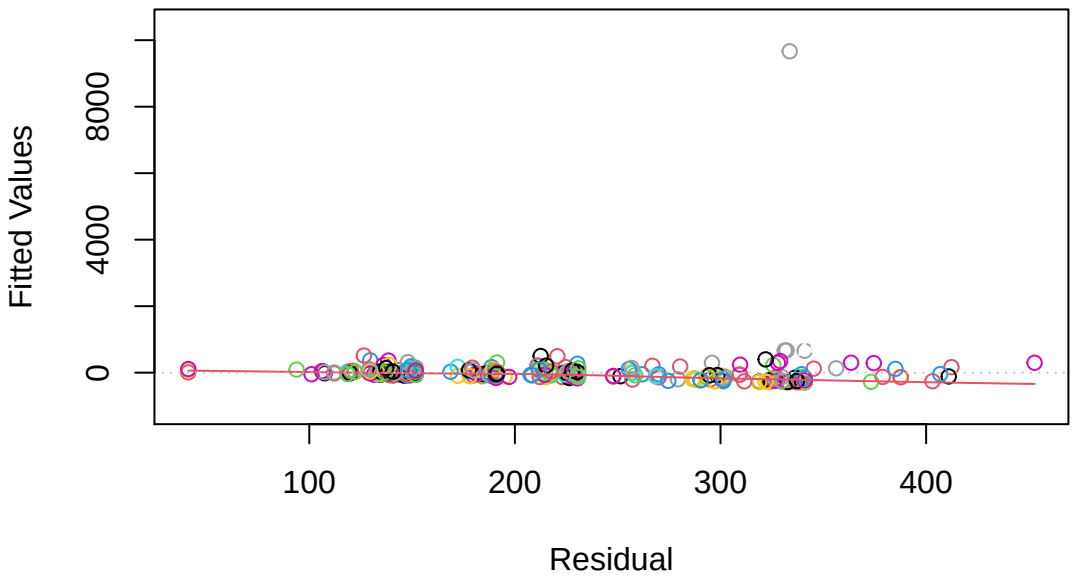


Table 4: ANCOVA Table before Any Removal

	Sum Sq	Df	F value	Pr(>F)
accommodates	0.1915194	1	26.6451148	0.0000006
as.factor(room_type_dummy)	0.2044173	1	28.4395300	0.0000003
review_scores_rating	0.1098683	1	15.2854127	0.0001251
accommodates:as.factor(room_type_dummy)	0.0005731	1	0.0797377	0.7779340
as.factor(room_type_dummy):review_scores_rating	0.0414009	1	5.7598944	0.0172774
Residuals	1.4950595	208	NA	NA