The price forecast for New York Airbnb

Team members:

1-Yu-Jui Chang,

2-Jingyi Xu

3-Wei -Chen Lu

4-Rongtao Ding

Executive Summary

Topic: The price forecast for New York Airbnb

We hope to provide price prediction service for users planning to go to New York and use Airbnb, and predict reasonable prices for them according to customers' requirements, so as to prevent them from choosing inappropriate houses and improve users' experience.

Data set

Our data set is from Kaggle- 2019 Airbnb listing information data set in New York City. We have 48560 observations in our data set, and the data has 16 variables in origin.This data set has perfect data, which enables us to understand the prices of different areas and room types in New York. We can use this data to predict Airbnb in New York.

Methods used

We hope to establish a number of different statistical models (linear regression, regression tree, k-nearest neighbors), and compare the advantages and disadvantages of each model to find out the most appropriate statistical model, and finally use it to predict the price.

Group’s findings

Although the housing price of Airbnb is cheaper than traditional hotels, there are some good and bad in it. Some unreasonable housing prices are easy to mislead consumers. Therefore, we hope to provide price reference for consumers and improve users' consumption experience through the establishment of models.

Introduction

With the development of the economy and the progress of transportation, tourism has become more and more an indispensable part of our life. It relaxes our mood and increases our knowledge, which makes many people fall in love with traveling.However, the high price of traditional hotels has brought a lot of troubles to many travelers. For many people, the biggest consumption during the journey is housing, while the appearance of Airbnb has brought people a new choice.Through B&B, they provide a platform for landlords to give people more accommodation options while traveling, but Airbnb also has some problems, especially some rooms that are not reasonably priced or unmanaged for a long time, which can make it difficult for users to choose.Through sorting and analyzing Airbnb data in New York, our team hopes to build a model to provide customers with certain price prediction and help them understand the reasonable price of the house in line with their expectations.This improves the experience of Airbnb users and reduces the impact of harmful information on users.We will use the methods we have learned this semester to analyze the data set we have chosen for Airbnb's renting situation in New York. We hope to understand the characteristics of Airbnb's houses in New York, and we can provide certain price guidance to those who plan to rent houses on Airbnb in the future through the establishment of models. Regression Tree, KNN, Forecast and other statistical methods were used to establish statistical models and compare their advantages and disadvantages, so as to finally select the best Regression model and make prediction.Specifically, people can decide what kind of room they want to live in according to their own budget and travel time, and our model can apply according to their preferences, so that he can easily find the room he might be interested in.The regression method used in this project lets the project team know the best decision to book Airbnb in New York.The project team hopes that users can use their budget and then add their preferences to locations and room types to find the best rooms on Airbnb.

Data

Our data set is from Kaggle- 2019 Airbnb listing information data set in New York City. We have 48560 observations in our data set, and the data has 16 variables in origin. We chose the neighborhood\_group, room\_type, price, minimun\_nights, number\_of\_reviews,

reviews\_per\_month variables for our final data set. Then we found there are some

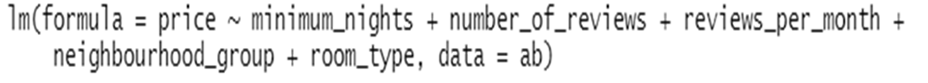
missing values in review\_per\_month, so we recode the NA value to 0. In

minimum\_nights, we thought our analysis is for a short trip traveler so we filter minimum\_nights less or equal to 14.

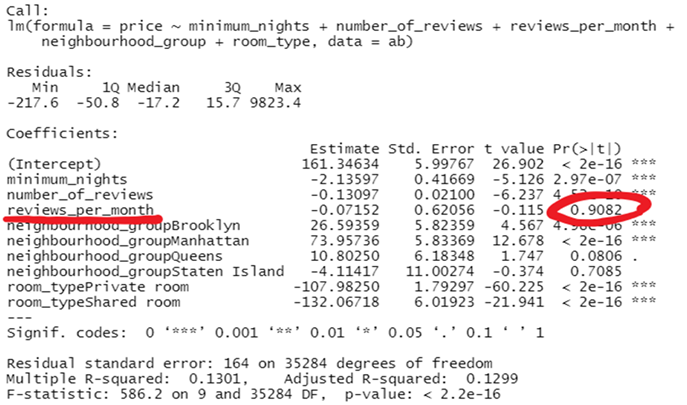
Analysis

Multiple Linear Regression

There are five variables used for this data analysis. The Price is the response variable. The independent variables are minimum\_nights, which we set equal and lower than fourteen days due to the consideration of travelers living in AirBnb, number\_of\_reviews, reviews\_per\_month, neighbourhood\_group, and room\_type. In addition,neighbourhood\_group, and room\_type are categorical variables in this data set, and the function of regression in R can automatically transform categorical variables into dummy variables. For this multiple linear regression model, our null hypothesis is that there is no linear relationship between the response variable (Price) and the independent variables. On the other hand, the alternative hypothesis is that at least one of the independent variables has linear association with the response (Price). The multiple linear regression formula as following:

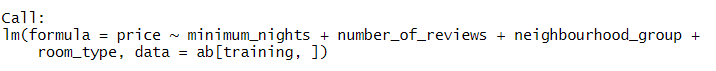


Here are the result of this regression:

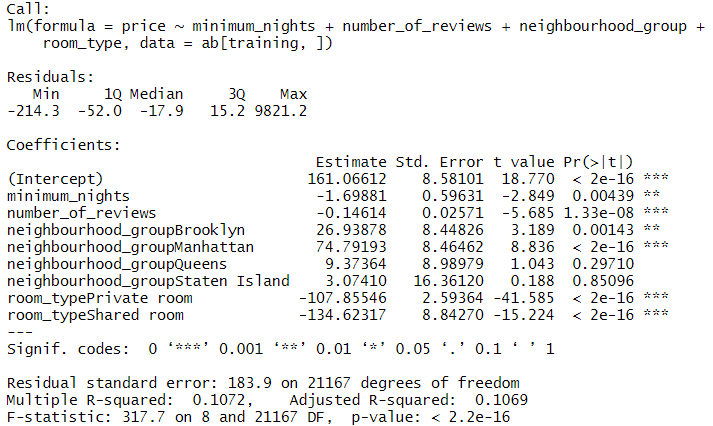


As we can see, the p value of the model < 0.05, the null hypothesis fails. There is enough evidence to support that at least one of the independent variables has linear association between the response(price) and the parameters at the significant level of 0.05. However, this result reveals that the p value of reviews\_per\_month is 0.9082, lower than 0.05, not significant with the response. Therefore, our team will create a new model without this parameter.

The new model’s formula as following:



The result of the new model as following:



The p value of the test is smaller than the significant level of 0.05, meaning that this model supports that there is at least one of the parameters that has linear relationship with the response (Price). However, we can find that the R-squared is only 0.1072, meaning that there is only 10.72% of the data points following this model. It’s no ideal. Also, the RMSE is 183.9. We will assess different models in the following method with the value of RMSE.

Regression Tree

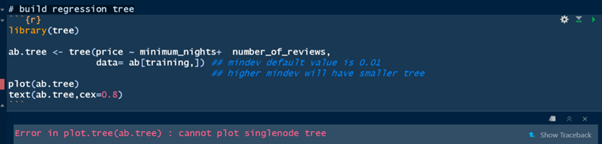
1. Split dataset into training and testing parts.

Use 60% data from our dataset to become the training part and the remaining 40% is testing part. Separate the result and others variables so that we can use it to test our model later.



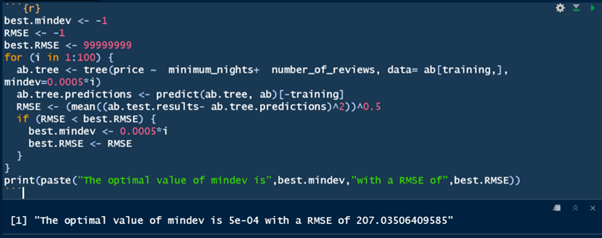
2. Build regression tree

Use a tree function to create a model. Price is our response variable and minimum\_nights and number\_of\_reviews are our predictors variables. Because we want to keep the model using the same variables, we also drop reviews\_per\_month here. However, we can find that the tree cannot be created at beginning. Because the default mindev is 0.001 and the higher mindev will have the smaller tree. So, we would try to find the best mindev for our regression tree.



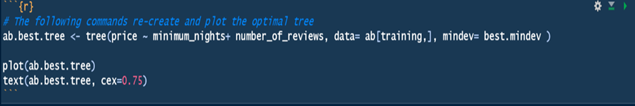
3. Find best mindev and create the best regression tree

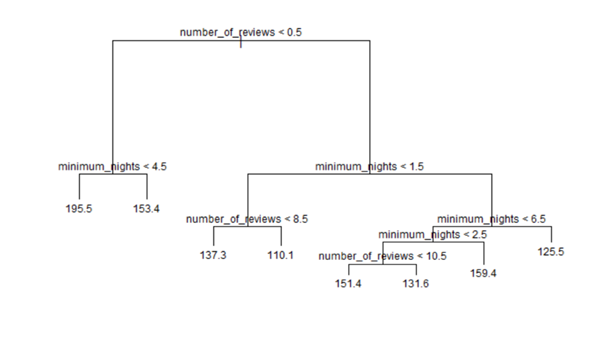
Use the for loop to find out the best mindev and create a regression tree. We find the optimal value of mindev is 0.0005 with a RMSE of 207.035, so we can use this mindev to create the regression tree.



4. Build the best regression tree

Use best mindev(0.0005) to create our regression tree.



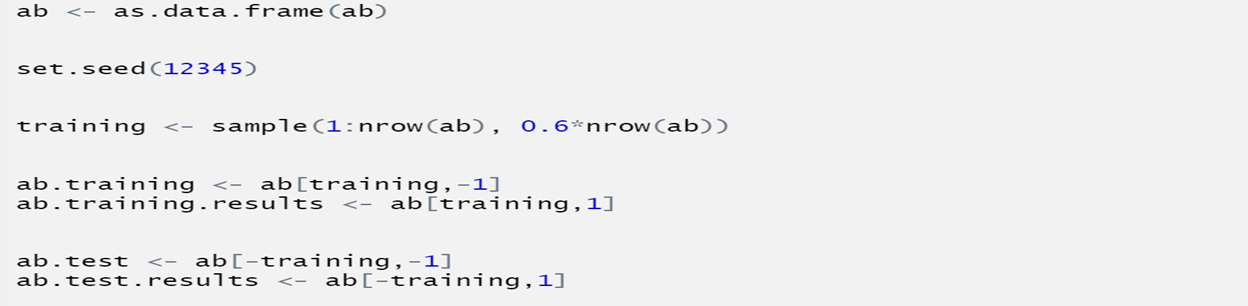


k Nearest-Neighbors (kNN)

The third method we used is k-nearest neighbors or named knn. We show the results by following three steps.

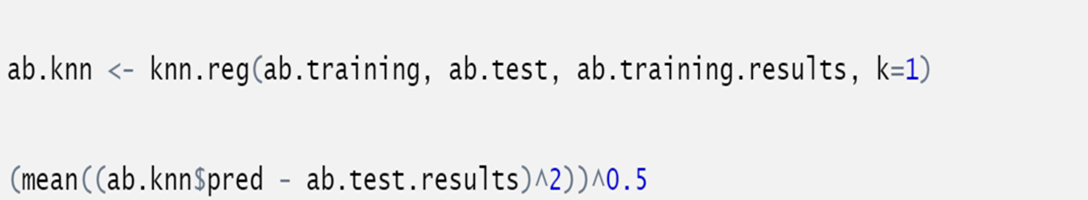
1. Split data to training and testing parts

In this step we do the same thing as Regression tree, randomly split 60% data to the training set and the remainder to the testing set. Then we split the first column of the dataframe to the result set. We stored them so that we can use them for the next step.



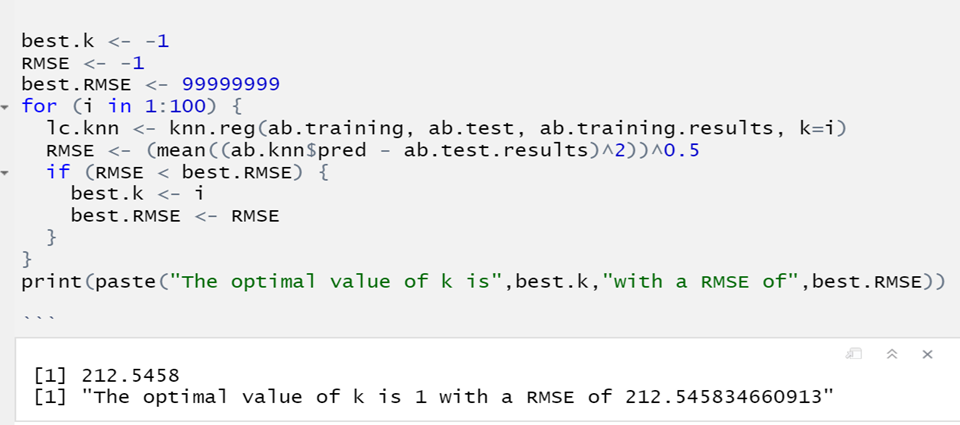
1. Build kNN model

Then we build the knn model by using the function knn.reg. We put the training set, testing set and result set into the function, the k value here is 1.



1. Find the best k value and root mean square error (RMSE)

We use the for loop to help us gain the best k value and RMSE. The result shows that the best k value is 1 and the root mean square error is 212.54.



Discussion and Conclusion

Contrast RMSE for the best method in this project：

Multiple Linear Regression: 189

Regression Tree：207.035

KNN：212.5458

189 < 207.035 < 212.5458

So we choose Linear Regression as our best model.

In conclusion, our team chose multiple linear regression as our best model for this project because it provides the smallest RMSE value, even though the R-squared is only 10.72 %. To sum up, our team concluded that this regression model can allow people who have interest in having an AirBnb in New York City to predict the room prices with specific conditions they made.

Besides, we found that Manhattan has the highest average price for three different types of room and Bronx is the lowest one among five neighborhoods. Since our data only have price, we cannot make further conclusions. If we want to know the reason for this situation, we should have more data about New York City, such as security or population dataset for further discussions.

Appendix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Average of price** | **Column Labels** |  |  |  |
| **Row Labels** | **Entire home/apt** | **Private room** | **Shared room** | **Grand Total** |
| Bronx | 127.5065963 | 66.78834356 | 59.8 | 87.49679193 |
| Brooklyn | 178.3381814 | 76.5081951 | 50.52784504 | 124.3965071 |
| Manhattan | 249.2574655 | 116.7766224 | 88.93319415 | 196.8789194 |
| Queens | 147.0319962 | 71.76245552 | 69.02020202 | 99.49399718 |
| Staten Island | 173.8465909 | 62.29255319 | 57.44444444 | 114.8123324 |
| **Grand Total** | **211.8066142** | **89.78702625** | **70.09318378** | **152.7250261** |

