Predicting Airbnb Listing Price In New York By Using Machine Learning Linear Regression

Nattawat Wattanawiput

October 10, 2021

1 Abstract

Airbnb has become increasingly popular among travellers for accommodation across the world. Accordingly, there are datasets being collected from the Airbnb listings with rich features. In this project, we aim to predict Airbnb listing price in New York City (NYC) with various machine learning in type of supervised learning in model of multiple linear regression, we have R-squared values of 0.4671 in train and 0.4614 in test on this dataset.

2 Introduction

Pricing a rental property on Airbnb is a challenging task for the owner as it determines the number of customers for the place. On the other hand, customers must evaluate an offered price with minimal knowledge of an optimal value for the property. This project aims to develop a reliable price prediction model using machine learning techniques to aid the property owners and the customers with price evaluation given minimal available information about the property. Features of Latitude, Longitude, Minimum nights, neighbourhood group, room type, availability 365 and customer reviews will comprise the predictors, and a range of methods from multi linear regression will be used for creating the prediction model.

3 Related Work

Predicting Airbnb Listing Price Across Different Cities. In 2019, Yuanhang Luo (royluo), Xuanyu Zhou (xuanyu98), Yulian Zhou (zhouyl).[1] Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. In 2020, Pouya et al. [2] Predicting Airbnb Prices with Machine Learning and Deep Learning by Nimon Dong et al. [3] Airbnb Price Prediction in the Age of Social Distancing by Richard Tran [4] New York AirBnB Regression Analysis, Visualization and Modelling, Price Prediction for AirBnb in New York Sayak et al. [4]

4 Dataset and Features

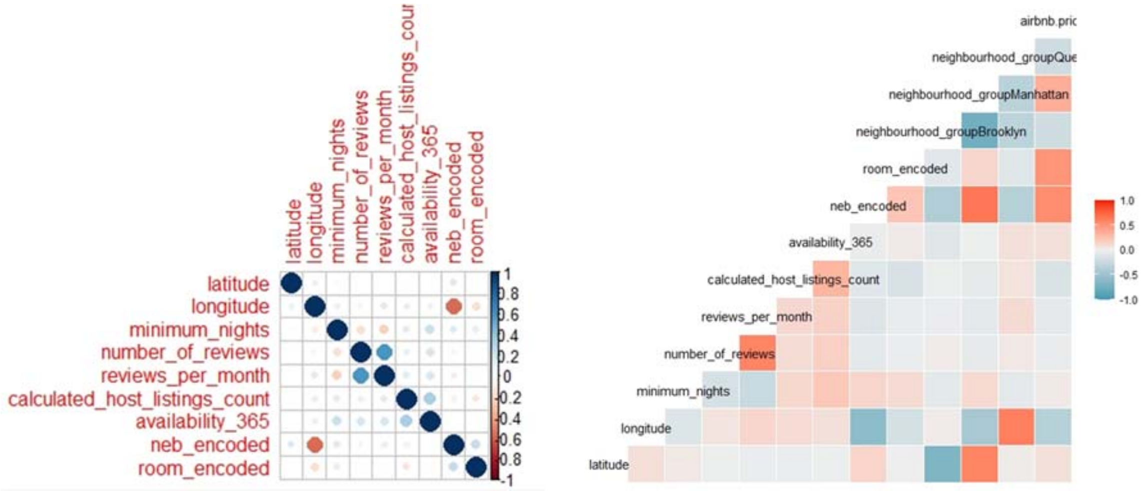
4.1 Dataset

We use Kaggle datasets New York City Airbnb Open Data. In dataset contain a detailed listing table with 16 raw input columns/features. The NYC dataset contains a total of 48895 listings and found 10052 missing values in reviews per month. We split the dataset into train: test with a ratio of 75: 25 for this dataset.

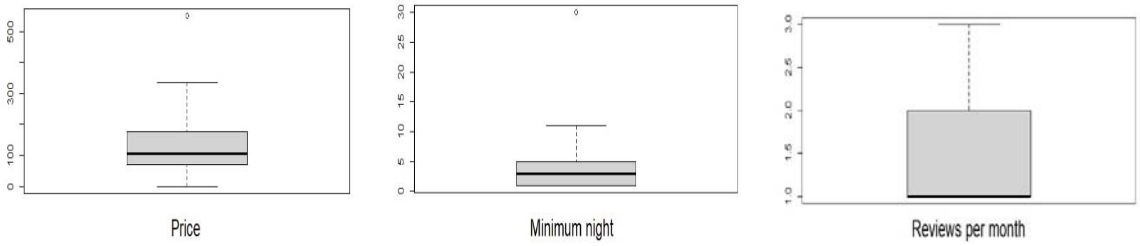
4.2 Features

Features were chosen only if they are informative and are likely to correlated with the target feature. Therefore, we eliminated features like id, name, host name and host id, which appear to be noise, features like last review that uncorrelation.

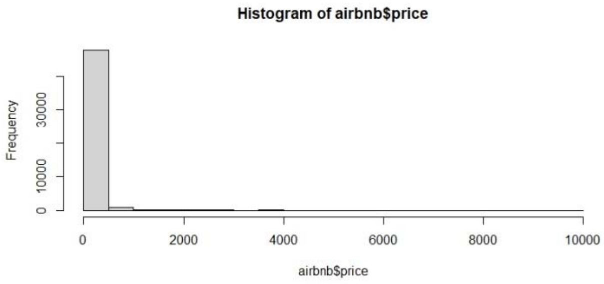
- **Target features**: For the listing price. As there exists abnormally high prices in the datasets, we have two activities were outlier removing and transformation to alleviate this problem. For data cleaning, we imputed data with price over 300 dollars per night by IQR method. we performed logarithmic transformation, and we found that logarithmic transformation works well for the price prediction and less error in results.
- **Numerical features**: For the numerical features in Latitude, Longitude, Minimum nights, Number of reviews, Reviews per month, Calculated host listings count and availability 365 we performed normalization transformation.
- **Categorical features**: For the categorical features, we transformed Neighbourhood feature by Bayesian encoders, Room type feature used OHE ordinal type and Neighbourhood group feature used OHE nominal type

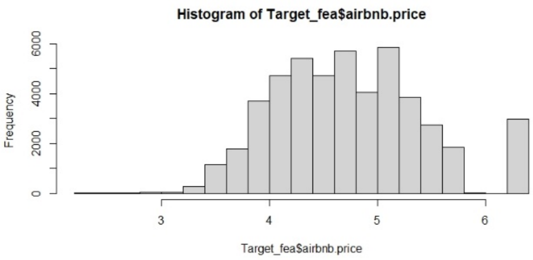


Correlation matrix



Outlier removing in numerical features





Before logarithmic transformation **After logarithmic transformation**

5 Methods

5.1 Multiple Linear Regression

Multiple linear regression has one y and two or more X variables. It is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable. Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.

Assumptions for Multiple Linear Regression:

1. A linear relationship should exist between the Target and predictor variables.

2. The regression residuals must be normally distributed.

3. MLR assumes little or no multicollinearity (correlation between the independent variable) in data.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,\ p-1} + \varepsilon_i$$

Y = Dependent variable / Target variable

β_0 = Intercept of the regression line

β_1, β_2, … β_n = Slope of the regression lime which tells whether the line is increasing or decreasing

X_1, X_2, … X_n = Independent variables / Predictor variables**

ε = Error

5.2 Evaluation metrics

Evaluation metrics for regression problems as below:

Coefficient of determination (R2 score) is used to evaluate the performance of a linear regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable(s). It is used to check how well-observed results are reproduced by the model, depending on the ratio of total deviation of results described by the model.

$$R2 = 1 - SSres/SStot$$

Mean Absolute Error (MAE): the mean of the absolute value of the errors is the easiest to

understand, because it's the average error.

$$\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE): the square root of the mean of the squared errors is even

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_j)^2}$$

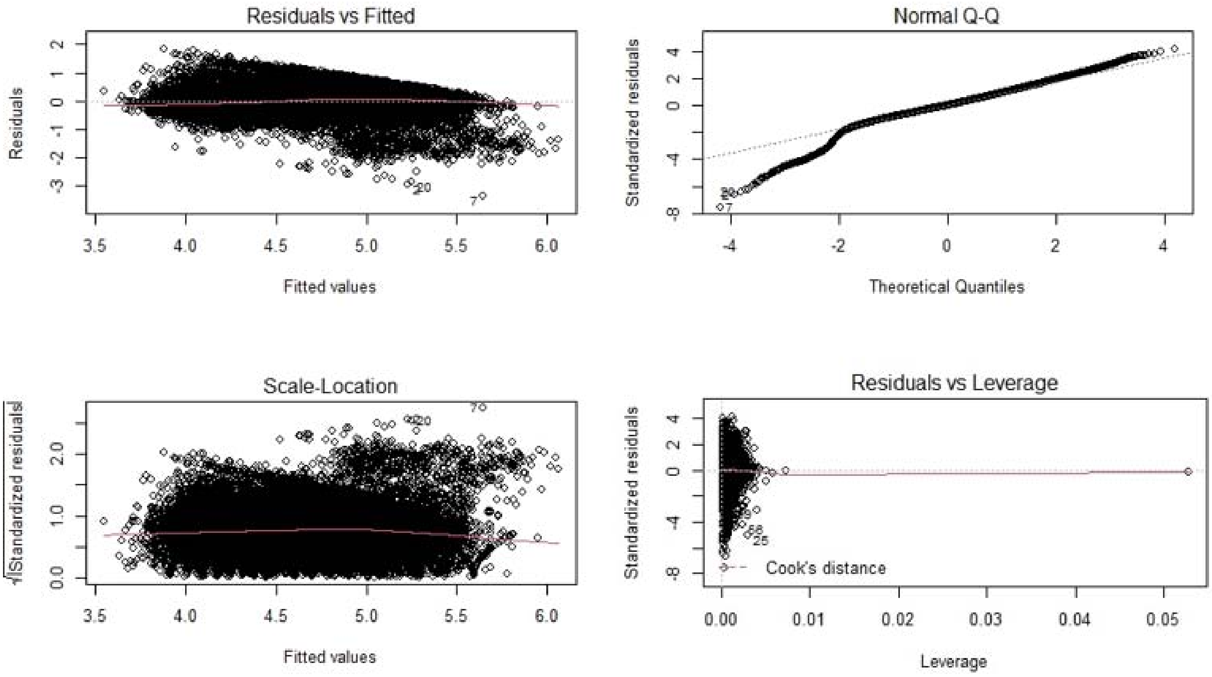
6 Results

6.1 Machine learning

For the results in this section, we used the default settings for machine learning approaches. we have experimented with different method to find the best score. First, we used to do the normal step of model multiple linear regression. Second, we tried to do process of variable selection in Backward stepwise selection. Third we checked and removed multicollinearity. The last, checked the assumption of the regression model for compare the best overall scores.

The linear model built by using original multiple linear regression has the best combination of RMSE and R-squared and is the smallest in AIC.

Testing Model Name	R-Squared(Train)	R-Squared(Test)	RMSE	MAE
Linear Regression	0.4671	0.4614	0.4445	0.3314
Backward Regression using AIC	0.4657	0.4618	0.4442	0.3315
Remove multicollinearity	0.4646	0.4589	0.4453	0.3321
Assumption of the Regression model	0.4644	0.4591	0.4453	0.3319

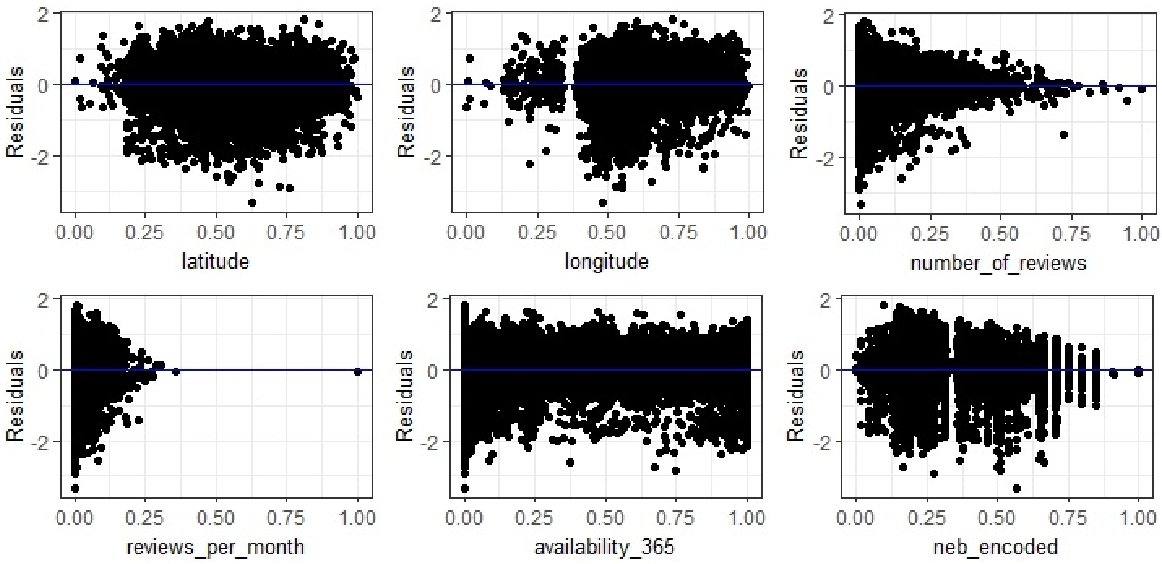


Model checking and diagnostics

7 Conclusion

This term project attempts to come up with the linear regression model for predicting the Airbnb prices that we used R programming for do the task of data science. Machine learning techniques of multiple linear regression with feature importance analyses and regression analysis that we have to find the best model with limit condition, employed to achieve the best results in terms of Root Mean Squared Error, Mean Absolute Error, and R2 score. Among the models tested to original multiple linear regression performed the best and produced an R2 score of 46.14% and a RMSE of 0.4445 on the test set. This level of accuracy is a not promising outcome given but we can use total process and method in regression analysis for Data science task in the real world.

Plots of Residuals vs Predictor Variables



Code

1.Data Pre-processing Airbnb Listing Price

https://drive.google.com/file/d/1Z1tSUtsM2p4BdqA_xLWvpoJmrviCHWTC/view?usp=sharing

2.Data Pre-processing Airbnb Listing Price file CSV

https://drive.google.com/file/d/1vMppKCY-TJ__dCse15u3frAEKaype9Th/view?usp=sharing

3. Predicting Airbnb Listing Price

https://drive.google.com/file/d/14_lRFAYCMTD5aNjk9f956Gu4iXK8TCdt/view?usp=sharing

References

[1] Peter Bruce and Andrew Bruce. Practical Statistics for Data Scientists. In 2017, Chapter 4 Regression and Prediction.

[2] Jiawei Han University of Illinois at Urbana-Champaign Micheline Kamber. Data Mining: Concepts and Techniques, 2006. Chapter 2 Data Pre-processing.

[3] Hadley Wickham and Garrett Grolemund. R for Data Science Import, Tidy, Transform, Visualize, and Model Data, 2017. Data Visualization with ggplot2, Part I. Explore.

[4] Yuanhang Luo (royluo), Xuanyu Zhou (xuanyu98), Yulian Zhou (zhouyl). Predicting Airbnb Listing Price Across Different Cities, December 14, 2019.

[5] Sayak Chakraborty | Samreen Zehra | Niharika Gupta | Rohit Thakur. New York AirBnB Regression Analysis, Visualization and Modelling, Price Prediction for AirBnbs in New York, 2019

[6] New York City Airbnb Open Data: https://www.kaggle.com/new-york-city-airbnb-open-data