2021

[PREDICTING THE PRICING OF AIRBNB LISTINGS]

1.1 Background

A dataset was posted on Kaggle with the following description:
'Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019. This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.'

Alongside this description was a dataset from Airbnb. One of the questions asked was: are we able to predict the prices of the listings based on the dataset provided.

1.2 Solution

To solve this problem a regression model will be created to try to predict the prices of the Airbnb listings in the dataset. As the dataset contains information about geographical location and we are able to supplement the dataset with data from Foursquare. We hypothesise that adding information about the nearby venues will give a better prediction compared to the original dataset.

1.3 Interest

The parties that may be interested in the results are the original poster of the dataset as well as Airbnb.

2. Data sources

Data about AirBnB locations and prices were found in a Kaggle dataset from here. The dataset contains information on the owner of the listing, name of the listing, the neighbourhood group, the neighbourhood, latitude and longitude, room type, availability, review metrics, minimum amount of nights, and price.

To complement this dataset Foursquare API was used to collect data on the venues surrounding the respective Airbnbs. The latitude and longitude of the Airbnbs was used to derive the nearby venues in a radius of 500 meter of the listing.

3.1 Methodology

Kaggle has been used to get the used dataset. Apart from this the dataset was supplemented using the Foursquare API. To map the different listings the geocoder API was used to get a clear map. As we are limited by the number of API calls we can make in a day, it was chosen to take the top 800 most popular listings. To reduce the bias of differences in neighbourhood group that were not based on the nearby venues it was chosen to take only the listings in Manhattan.

For machine learning techniques we will be using ordinary least squared linear regression to predict the pricing for different listings. To make a comparison between the supplemented dataset and an empty dataset, two different models will be build using the same technique. One will be the build using the supplemented dataset containing 800 different listings in Manhattan, the other will be build using the original dataset and also be using the same listings. Also a decision tree will be build for the supplemented dataset to see if this leads to better results.

The dataset will be split using a 80/20 split, with 80% for training and 20% for testing. One-hot encoding will be applied the Foursquare data as well as other nominal categorical variables. Data will be transformed to have a mean of 0 and a standard deviation of 1. Due to time constraints backwards selection was not used in the feature selection. Instead the dataset was supplemented with the most common venues while keeping a 10:1 ratio for venues to parameters to reduce the risk of overfitting. If improvement is seen using this amount of data, it is suggested that using the total dataset and all parameters will also lead to the same or better improvement.

The model performance will be evaluated using R^2. The model will the highest R^2 will be the best model.

3.2 Data cleaning

After selecting the available listings the dataset was searched for missing values. Rows containing unnecessary columns were dropped such as information about the host, the timing of the last review, and the ID of the listing. Afterwards the dataset was checked for missing values and outliers. There were no missing values in the dataset. The price column seemed to have a few outliers ranging up to a price of over $10.000 dollars per night. As less than 1% of the dataset had a price of over $1000 dollars per night, these were considered as outliers and were therefore dropped from the dataset. Listings with a price of 0 were also dropped, as these apartments are currently unavailable and would skew the results towards zero. No other outliers were found.

3.3 Transforming data

The price of different listings was skewed towards zero. To combat this a log transformation was used to get the price more normally distributed. The same technique was tried on the minimum amount of nights. The skewness was so big that even log transformation did little to nothing. Therefore it was chosen to transform the parameters from continuous to two categories, less than 6 days and more than 6 days.

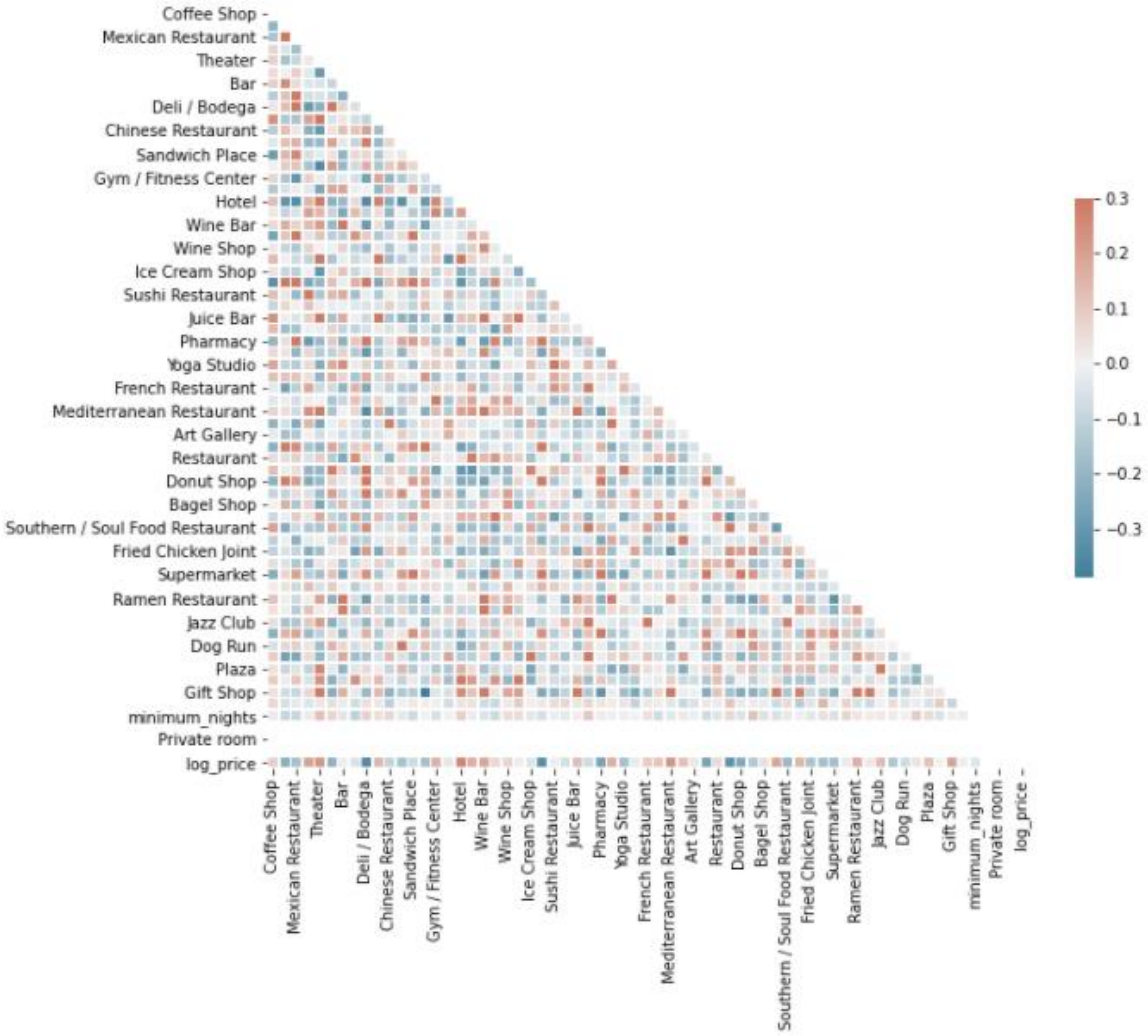
4. Results

Based on the number of reviews per listing, the 800 most popular were selected. Using the geocoder API these were plotted on the map below.

The parameters used from the original dataset were the minimum amount of nights, the availability, and of course the outcome price. The dataset was supplemented by 59 more parameters, which were the venues from Foursquare. Before further modelling the parameters were checked for

multicollinearity by creating a correlation matrix. Correlations greater than 0.8 were dropped from the dataset.

Figure 1. Correlation Matrix



Three models were fitted, two on the supplemented dataset and one on the original dataset. The results are shown in table 1 below. The calibration of the predictions is shown in figure 2:4.

Table 1. R^2 for the different models.

	OLS (supplemented)	Decision Tree	OLS (original)
R^2	**0.178**	**0.233**	**0.010**

Figure 2. Linear regression on the supplemented dataset, predicted against actual values.

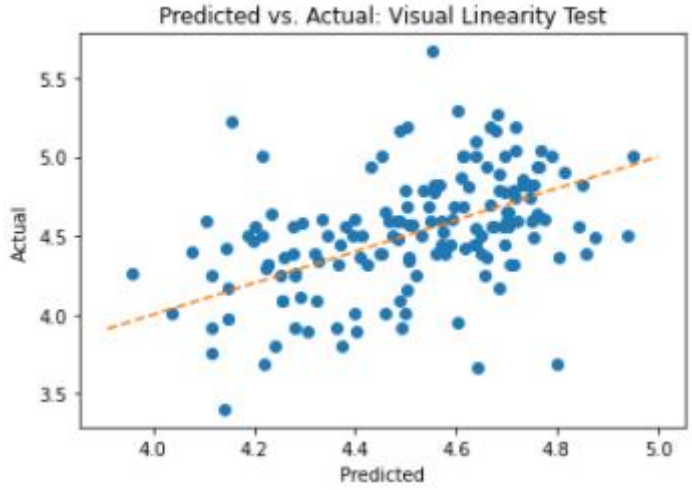


Figure 3. Decision tree on the supplemented dataset, predicted against actual values.

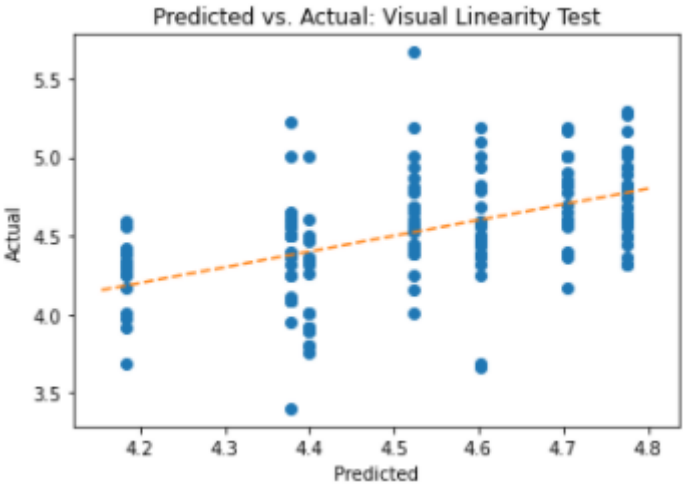
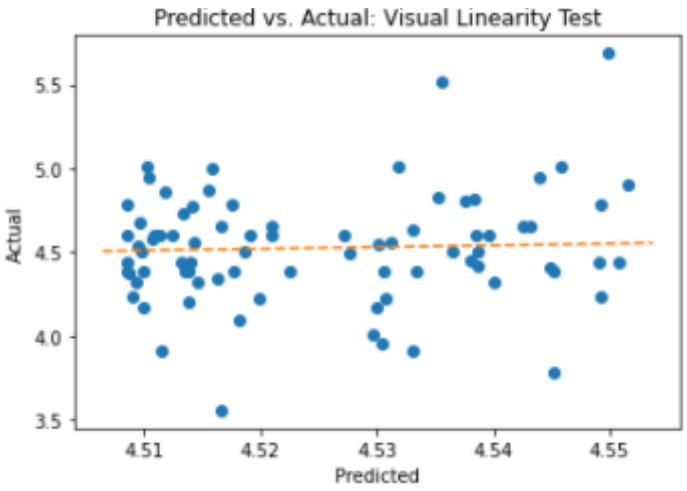


Figure 4. Decision tree on the supplemented dataset, predicted against actual values.



5. Discussion

Based on the R^2 it would suggest that the supplementation of additional data using the Foursquare API is beneficial in improving the predictions on price using the Airbnb dataset. It must be said that the prediction using OLS on the original dataset was very poor. This is likely due to the lack of appropriate information relevant for prediction. To improve predictive ability altogether the data should be supplemented with information on for example the size of the listing, as well as how many beds are available. This is likely to be more informative for predictions than the current information.

The poor results may also be partly attributable to the lack of sample size and thus the reason for dropping many of the available features. This has led to immense information loss. But because I can only make so many API calls a day the available data was cut short. If the entire dataset could be used including all the different features it is likely that the model will perform better than suggested in this paper. The lack of backward selection also hindered the model building. This is however beyond the scope of this course. This would however help improve model performance by including the relevant features for prediction instead of the most common. This would likely explain the low performance on the supplemented dataset.

It must also be said that OLS may not be fitting in predicting the price, as one of its assumptions may be violated. I have spent time checking assumptions of linearity, homoskedacity, and multicollinearity. However I ran out of time when trying to create a Q-Q plot to check for the normality assumption. This should be tested before making any other conclusions based on the results. The assumption of independency of samples may also be violated, as information on one listing with the nearby venues gives information on the nearby venues of other listings in it vicinity. Due to this reason a multilevel analysis may be more appropriate. If a multilevel analysis is considered for further modelling, it should also be considered to take into account the different boroughs as information other than the nearby venues could be valuable in predicting pricing. These analyses however are beyond the scope of this project.

6. Conclusion

Based on the results of the different models we can conclude that the addition of the Foursquare API leads to better predictions compared to the original dataset from Airbnb. The predictions are however poor at best and should therefore not be used to make any inferences. It is recommended that if predicting the pricing of different listings is a priority to supplement the dataset with other relevant information apart from geographical location.

Appendix.

Figure 6. Residual of linear regression (supplemented dataset).

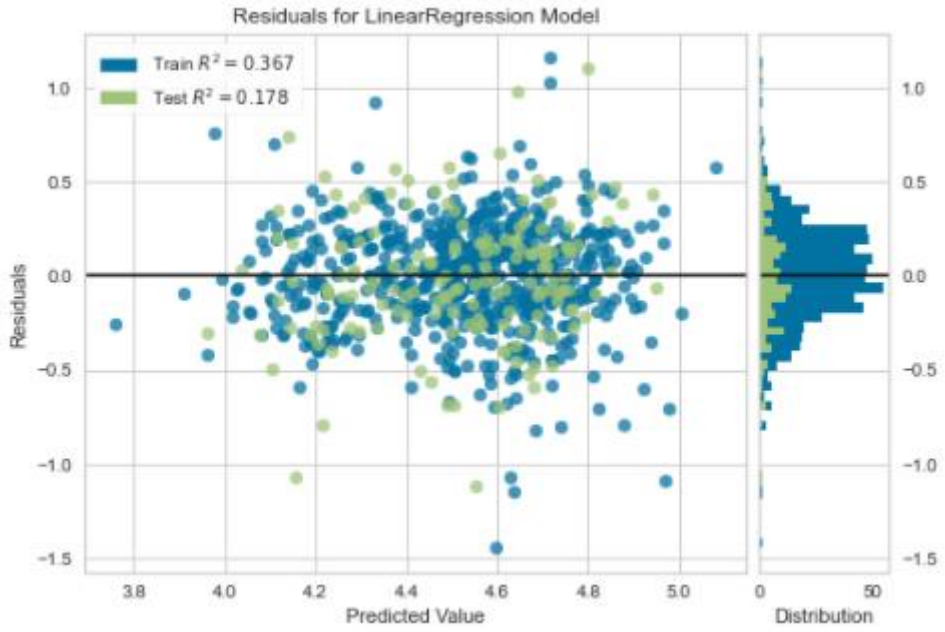


Figure 7. Residual of linear regression (original dataset).