New York Airbnb Price Prediction

Ye Jin

17 Jan 2020

1. Introduction

1.1 Background

Airbnb is a rental platform which offers rental properties, homestay and paying guest options along with tie-ups with restaurants and resorts to provide a variety of shared and independent short-term need based options to tourists, students and business professionals. The business model for the company includes commission-based economy and it depends on the type of rental services that the company provides and how better the accommodation facilities are. It is an online rental platform and thriving a lot with the advancement of technology.

1.2 Problem

However, with the advancement of technology, several emerging companies join the league of the online rental platform and giving hard competitions to Airbnb. A few of these companies are HomeToGo, Tripping.com, HomeAway, FlipKey, which has emerged in a shared rental market with a similar concept to Airbnb. In order to take a place in the market, Airbnb is trying to introduce and price its services based on the features listed for the lodging so that they can provide customized solutions on the basis of the customer's budget and demand.

1.3 Interest

As a data consulting firm, we were given a task to design a model based on statistical learning techniques to predict nightly pricing for their listings with a focus on properties based on New York, US. For the purpose of developing a best statistical model for this case, models like multiple linear regression, ridge, lasso, Bagging and Random forest were used. Using the RMSE value as the criterion to decide which is the best prediction model. The lower the RMSE value, the better the prediction model.

2. Data processing

2.1 Data source

A well-developed dataset could be found from Kaggle dataset “New York City Airbnb Open Data”. It has 16 columns including room features, locations as well as reviews. The dataset includes various types of data in numeric and categorical data and, also includes missing values which needs to be fixed to make future predictions.

2.2 Feature Selection and fix missing values

```
id                                  0
name                               16
host_id                             0
host_name                          21
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                     10052
reviews_per_month               10052
calculated_host_listings_count      0
availability_365                    0
dtype: int64
```

In this dataset, there are only several columns have missing values, name, host_name, last_review and review_per_month. We will drop the name, host_name and last_review as they are irrelevant to our analysis. To fix the missing value in review per month, we can simply replace the missing value by 0 as this will not affect our analysis as well. After processing, there are no missing values.

```
host_id                         0
neighbourhood_group             0
neighbourhood                   0
latitude                        0
longitude                       0
room_type                       0
price                           0
minimum_nights                  0
number_of_reviews               0
reviews_per_month               0
calculated_host_listings_count  0
availability_365                0
dtype: int64
```

3. Exploratory Data Analysis

Exploratory data analysis is a procedure of explaining the data using data visualization techniques, like graphs, plots to understand the characteristics of datasets. This method helps in understanding the relationship between different parameters and their importance in predicting the response variables. Few of the methods that we have used in our report are histogram, scatterplots, box plot, heatmap, run chart etc.

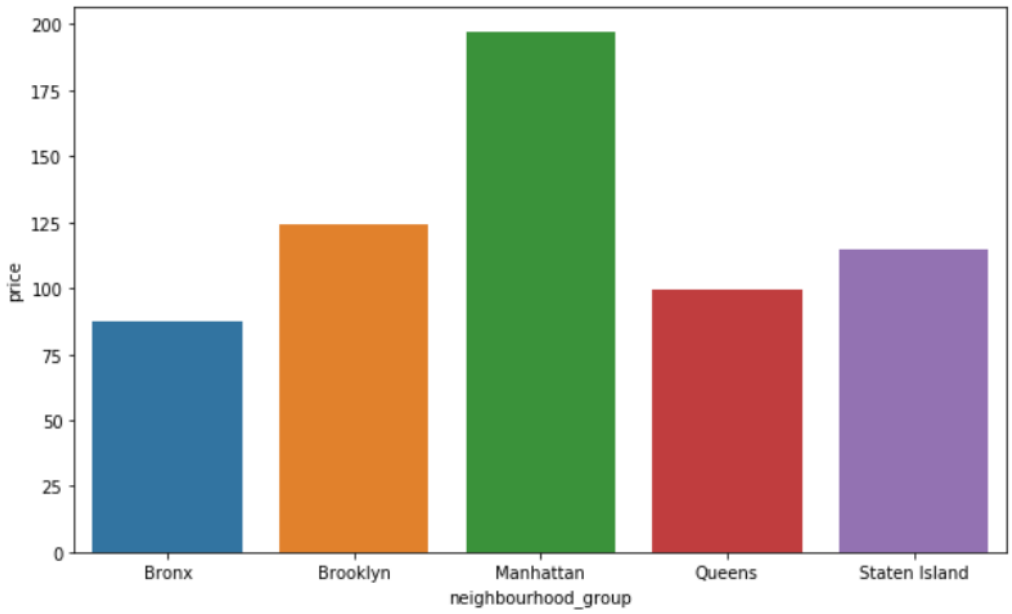
3.1 Correlation Matrix heatmap

Correlation matrix as heatmap is used to check the correlation among different numeric variables as there are multiple columns in the dataset.

Correlation among all the variables was plotted in the form of the heat map. Here, light orange means positive, and dark means a negative correlation. Higher the value correspond to colour is scale, grander the magnitude of correlation it represents. The one with shades of orange are highly correlated and have strong relations. It was observed that review per month and number of review have the highest correlation Metrix, and there are chances of multicollinearity. To avoid the case of multicollinearity, we have further used models like lasso and ridge.

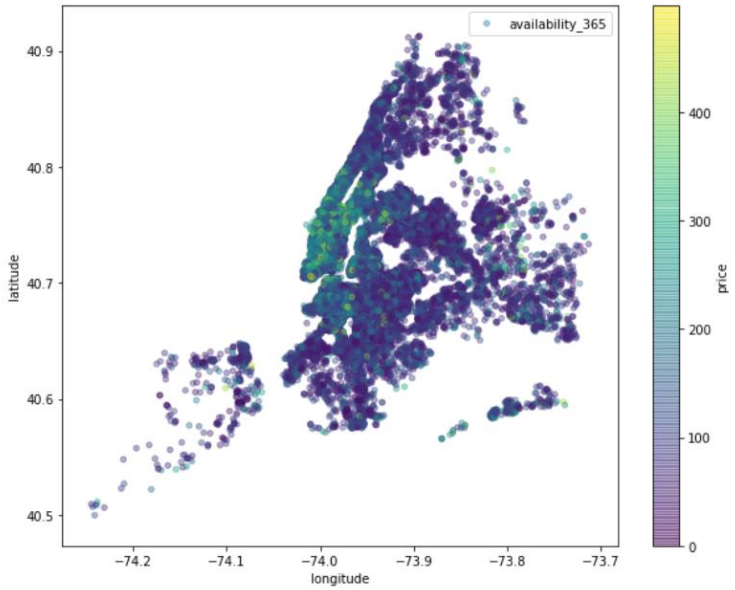
3.2 Relationships between neighbourhood and price

Location is one of the most significant factors affecting house price, which applies to Airbnb as well.



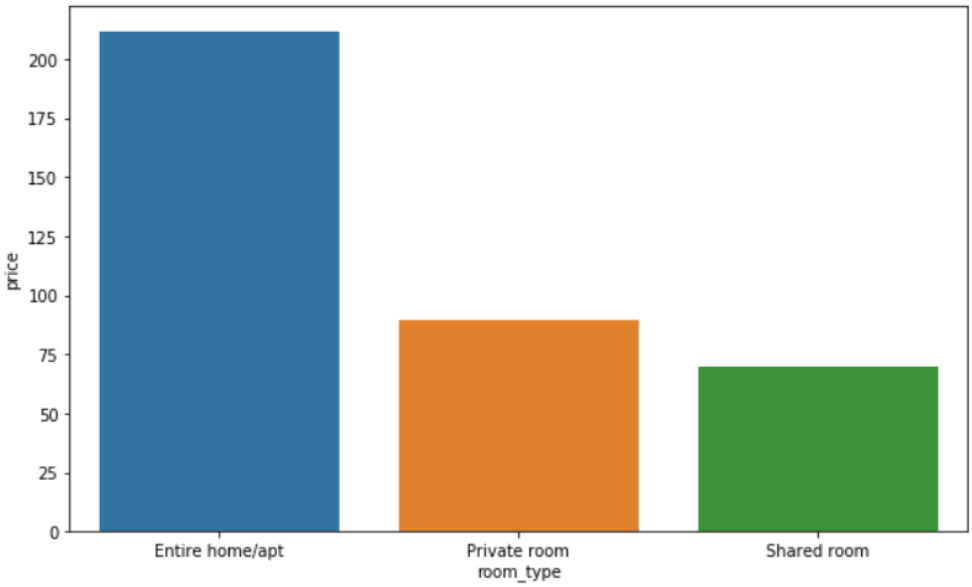
We took the average price of each neighbourhood group and plot this bar chart. It is not surprising that Manhattan has the highest Airbnb price as Manhattan is known as one of the most expensive area in all ways. It followed by Brooklyn, Staten Island, Bronx and Queens.

The price distribution can be visualised by a map as following:



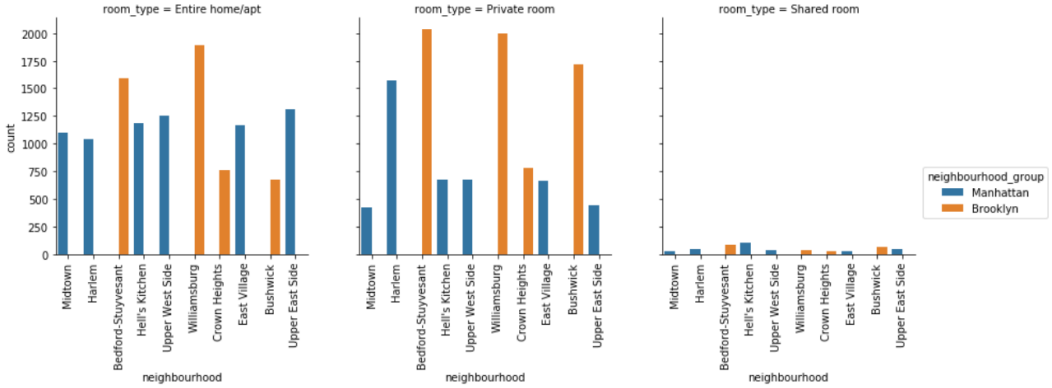
In the above map, the darker the dot, the lower the price.

3.3 Relationship between room type and price



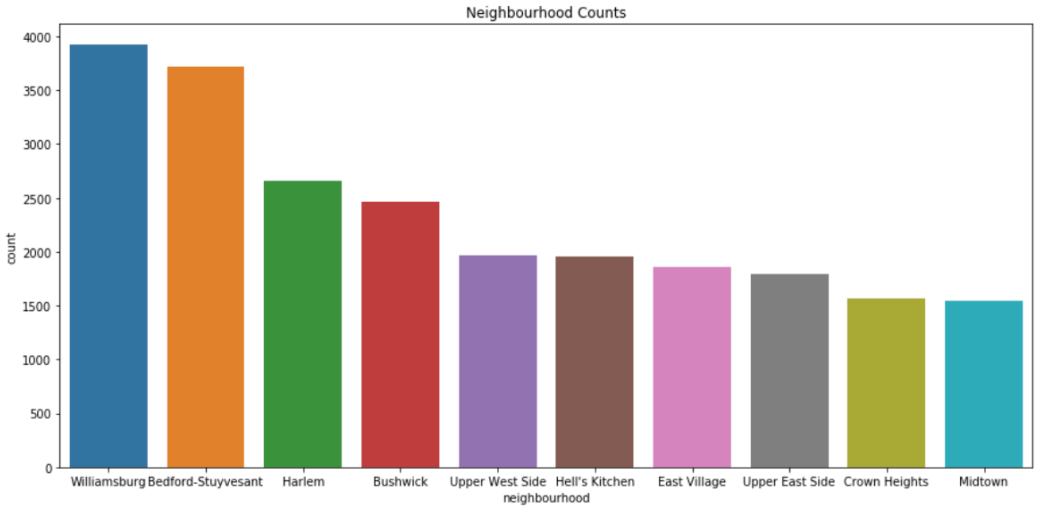
We did the same to analyse the relationship between room type and price. As shown in the above bar chart, the price of entire home has a higher price, followed by private room and shared room, which is true in real cases.

We also investigated which type of property is offered the most in the top 10 neighbourhoods in Manhattan and Brooklyn.



It is obvious that there are more entire home/apt and private room than shared room, indicating that entire home and private room are more popular and have a higher chance to be chose.

3.4 Top neighbourhoods and top hosts



The above bar chart shows the ten neighbourhoods offering the most Airbnb Property. These neighbourhoods might be hot tourist destinations, with large potential customers.

4. Methodology

4.1 Multiple linear regression

Multiple linear regression analysis is widely used in supervised statistical learning, to evaluate how an outcome or response variable is related to a set of predictors. The advantage of using MLR here over other generalisers is its interpretability as the price generated by it indicate the different contributions that each feature makes for class prediction.

Admittedly, multiple linear regression model does have some insufficiency which affects prediction accuracy and interpretability. For instance, the fitting procedure, Ordinary Least Squares, tends to perform poorer as the number of predictors approaches substantially large, leading to overfitting outcomes. Moreover, multicollinearity problem can also affect the multiple regression model by leading to increased standard errors of the coefficients, hence makes some variables statistically insignificant when they should be significant. Based on these, though it could be slightly improved by adding an additional variable to rise the variance, and by remove highly correlated predictors from the mode to minimize multicollinearity, MLR is no longer hold its advantage considering its low effectiveness and limit performance improvement due to the great number of explanatory variables about Airbnb prices.

4.2 Lasso

Lasso regression performs L1 regularization, which also add a penalty that equal to the absolute value of the magnitude of coefficients to the normal OLS model. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model while the ridge regression doesn't result any elimination, this could makes the Lasso tend to perform better than the Ridge.

Different to ridge regression, lasso regression can drive some coefficients to zero given a suitable lambda value. The larger the value of lambda the more features are shrunk to zero. This can eliminate some features entirely and help us simplifies the predictors, thus reduce multicollinearity and model complexity. Predictors not shrunk towards zero signify that they are important and thus Lasso model not only helps in reducing over-fitting, but it also allows feature selection.

The main problem with lasso regression is when we have variables correlated with each other, it will retain only one variable and eliminates other correlated variables by setting them to zero. That will possibly lead to some loss of information resulting in lower accuracy in our model.

4.3 Ridge

Shrinkage methods are one of the potential approaches dealing with overfitting and model with a too high degree of complexity. By scarifying some bias, shrinking the coefficients leads to a lower variance and in turn a lower error value. One of the best-known approaches is the ridge regression.

Ridge regression uses L2 regularization which adds the penalty term to the OLS equation. The L2 term is equal to the square of the magnitude of the coefficients. In this case, a zero lambda(λ) means the basic OLS equation, but if it is greater than zero then a constraint is added to the coefficients which

results in minimized coefficients that trend towards zero the larger the value of lambda. λ becomes larger, the variance decreases, and the bias increases. Thus, by modifying the λ, an optimal trade-off judged between bias and variance can be achieved.

Ridge regression prevent multicollinearity and decreases the complexity of a model but does not reduce the number of variables, it rather just shrinks their effect, hence decrease the instability of the model.

4.4 Elastic Net

Another commonly used model of regression is the Elastic Net which incorporates penalties from both L1 and L2 regularization, in which keeps all predictors but narrow the coefficient to solve related problems.

In addition to setting and choosing a lambda value, elastic net also allows us to modify the alpha parameter where $\boldsymbol{\alpha} = 0$ corresponds to ridge and $\boldsymbol{\alpha} = 1$ to lasso. In detail, if you plug in 0 for alpha, the penalty function reduces to the L1 (ridge) term and if we set alpha to 1 we get the L2 (lasso) term. Therefore, we can choose an alpha value between 0 and 1 to optimize the elastic net. If the model is effective, it could reach a balance between shrinking and eliminating. Overall, elastic net gets the advantages and eliminates some disadvantages of ridge regression and lasso.

4.5 Bagging

Bagging, also known as bootstrap aggregation, is a general-purpose procedure for reducing the variance of a statistical learning method. In practice, we take many training sets from the population and build a separate prediction model using each training set. Then we average the resulting predictions. Consider bagging for regression trees, we construct B regression trees and using B bootstrapped training sets, then average the resulting predictions.

Bagging helps reduce variance and thus helps us avoid overfitting (Kandan, 2018). Especially, the bagging is able to deal on the data with large size and high dimensions, which could predict prices easily with variety of predictors. However, there is loss of interpretability of the model. The final bagged classifier is not a tree, and so we forfeit the clear interpretative ability of a classification tree thus can possibly be a problem of high bias if not modeled properly. Another main disadvantage is that while bagging gives us more accuracy, it tends to be more expensive and may not be desirable on some cases (Kandan, 2018).

4.6 Random Forest

Random Forest is an extension over bagging. It takes one extra step where in addition to taking the random subset of data, it also takes the random selection of features rather than using all features to grow trees (Nagpal, 2017). Random Forest also help overfitting most of the time, by creating random subsets of the features and building smaller trees using these subsets. The main limitation of Random Forest is the ineffective real-time predictions due to large number of trees. Even a more accurate prediction requires more trees, it also become slower which not suitable when run-time performance is important in the real-world cases (Donges, 2018).

5. Results

Method	RMSE	R2 score
MLR	176.58337	0.07480
Lasso	176.58618	0.07477
Ridge	176.57157	0.07493
Elastic Net	176.57958	0.074843
Bagging	178.10512	0.058788
Random Forest	161.56591	0.22547

In this analysis, we use RMSE and R2 score to evaluate the models.

As shown in the result table, random forest has the lowest RMSE and highest R2 score, indicating it outperforms other models. For MLR, Lasso, Ridge and Elastic Net, there are no significant difference in RMSE and R2 scores so they might not be the suitable model here. Bagging has the worst performance as it has the highest RMSE and lowest R2 score. Therefore, the best model we will choose after this analysis is random forest.

6. Discussion and conclusion

Although we use six methods in this analysis, there are still many limitations that will affect the accuracy.

Firstly, there might be other models that could produce more accurate predictions. Each model has its constrictions and that is the reason we choose multiple models in the analysis. However, due to the limitation of our knowledge, we are unable to cover all models.

Secondly, the price of an Airbnb property might be affected by other features, for example, the internal design. Personal preference takes a big role in property selection and there are always trends in interior design style. For properties with more popular style design, the price might be higher. This hypothesis should be further analyzed with more data.

In conclusion, this analysis is aiming to make predictions on Airbnb prices in New York, based on property features. This prediction can be helpful for people who would like to start a Airbnb business or tourists who are looking for an Airbnb. We cleaned and explored the data, investigated significant relationship between features. Six different methodologies are used to train and test models. At last, we found that random forest could be the model to give most accurate prediction.