Data Analysis and Prediction on Airbnb Listings in Seattle, WA, United States

Li, Tianyi[1], Nie, Claire[2], and Sabbineni, SreeSindhu[3]

[1,2,3]University of Washington, Tacoma
[1]Student ID: 1827924
[2]Student ID: 1466737
[3]Student ID: 1875305

December 5th, 2018

Abstract

In the center of this project, we build a prediction model to predict the prices of Airbnb accommodations in various neighborhoods of Seattle, WA. We also suggest the most popular places to stay generally based on the number of reviews in a neighborhood, and the average number of days the listings in the neighborhood are booked in the next year. We have used the data of Airbnb listings in Seattle area collected from 'Insideairbnb' and performed data cleaning and data wrangling as per our requirements. We have applied a number of regression algorithms for price prediction and linear regression produced the lowest RMSE value of 0.312 ,which we considered as a good model. Popularity is determined from average reviews per month for a listing and number of days it is booked for next year. We have programmed our model in Python and used Pandas, Numpy, Matplotlib, Seaborn and Sklearn libraries.

1 Introduction

Airbnb is an increasingly popular option for booking accommodations among travellers from all over the world tend to use it for their accommodations. Suppose someone who has never been to the Seattle area is planning a week-long trip to this amazing city and wishes to use Airbnb for accommodation. This visitor would like to know what are the most popular neighborhoods and what neighborhood to stay at based on their budget. Our price estimator is a useful tool for the travellers to make accommodation decisions.

Imagine someone is looking to purchase an investment property in Seattle and plans rent it on Airbnb to make a profit. Which neighborhood should the investor look to purchase based on his or her expected rental income? Or,

consider local landlords who have empty space in their current homes and are looking to make some extra income from Airbnb. How much potential income they could generate from renting their space on Airbnb? Would it be worth the trouble? Our price estimator would show the most popular neighborhoods and the rental prices in order to help investors and existing homeowners make the best investment decisions.

Solving the proposed problem can help travellers plan their trips and help investors make better investment decisions by estimating potential income from Airbnb rentals.

2 Related work

We investigated two related works. The first is a series of four articles called "Making Models Airbnb Price Prediction: Data Analysis" written by Philip Mohun[2]. In his study, 74000 records of Airbnb data encompassing various U.S. cities. The author used log_price as target variable for this analysis. He produced a correlation matrix to visualize the multiple variable interactions. Through this multivariate analysis, the author identified features accommodates, bathrooms, and cleaning_fee to be most highly correlated to the target variable log_price. Mohun proceeds to transforming each feature into different forms to facilitate model fitting. After deciding on the most appropriate transformation for each feature, the author opted for generalized linear model with RMSE of 0.356.

Another research project on predicting listing price of Airbnb listings using Zillow Home Value Index (ZHVI) was done by Nicole Samrao[3]. The study covered 12 major U.S cities. To determine the most popular cities to stay, she first analyzed the average number of days the listings of a city are booked in the next year(365 days). The higher the number of days booked, the more popular the city. She then used the average number of reviews per city as another indicator of degree of popularity of a city since the more frequently a listing is booked, the more reviews it would have. Samrao then imported the ZHVI data from the Zillow website. the ZHVI contains the median home price for every city as well as the the 5 year change and the 10 year change as a measure of home price growth rate. The correlation analysis indeed shows a strong relationship between the median home price of a city and the price of an Airbnb listing. To build the linear regression model to predict price of a listing based on Airbnb dataset feature and the ZHVI, Samrao split the data with 80% as training data and 20% as test data to evaluate the model's performance. Samrao uses The R-squared coefficient of determination to measure the accuracy of the linear model in predicting the price of a listing from the selected features. The overall accuracy of the linear model was 56%. To improve the model, Samrao repeated the model fitting process by segmenting the dataset into two subsets: highly popular cities and cities that are not as popular. The model improved to 64% accuracy. Samrao then used Random Forest Regression on highly popular cities, the resulting R-square value is 71.2%, which significantly improved.

The main difference between the related work and our proposed problem is that the related work has a much greater scope covering multiple cities, whereas our problem concerns only neighborhoods in Seattle. We will not be using Zillow's Home Value Index, as this data is only for entire homes. Our dataset contains a significant number of private room and shared rooms.

3 Dataset

We have taken two datasets from the website called Inside Airbnb[1] which is an independent and non-commercial set of tools and data that allows us to explore how Airbnb is really being used in the cities around the world. As the scope of our project is small, we have considered only the datasets related to the Seattle area. For our datasets we selected 'listings.csv' file which contains detailed information and metrics for all listings in Seattle and 'reviews.csv' file which contains review dates for different listings in Seattle. In our datasets, 'listings.csv' contains 8635 records and 'reviews.csv' contains 356,804 records. All the data in each column of these two files are at the same level of granularity; each record is for a listing. Both files are in CSV and are rectangular tabular data in standard format. In each file, the columns hold different data types. Also in both files the data is flat. These two data set have listing ID as common label, which will be used to join these two datasets. In 'listings.csv', every record is unique and can be accessed by its primary key 'listing_id'. However, in 'reviews.csv' file, 'listing_id' and 'date' together are the primary key .

3.1 Data Preprocessing

3.1.1 Selecting Columns

'listings.csv' has the following columns: 'id', 'name', 'host_id, 'host_name', 'neighbourhood_group', 'neighbourhood', 'latitude', 'longtitude', 'room_type', 'price', 'minimum_nights','number_of_reviews', 'last_review', 'reviews_per_month', 'calculated_host_listings_count', 'availability_365'. Among all of these columns, for our prediction model, we only keep 'id', 'neighbourhood_group', 'room_type', 'price'. We renamed 'id' as 'listing_id' to facilitate merging with 'reviews.csv' later on. We will add a new column "avg_reviews_per_month' as described in the following subsections. 'reviews.csv' contains only two columns: 'listing_id' and 'date'.

3.1.2 Data cleaning

We restructured and modified our datasets as per our requirements. In 'listings.csv' a few records have price of 0, which we removed since these do not make sense. It is important to remove outliers in the data. To do so we calculated the mean price and standard deviation for each room type in each neighborhood. We remove listings with prices that falls outside 2 standard deviations from the mean price for each room type and each neighborhood. Although we will adjust

the number of standard deviations used in the Experiments section based on model performance.

3.1.3 Data Manipulation

The data in "reviews.csv" range from September 2008 to September 2018. To ensure timeliness, we will only use records with review dates in 2017 and 2018. In "listings.csv", the prices are the current listing price, therefore if a listing has had a review in 2017 or 2018, it means the price is representative of a fair market price. The last review column of "listings.csv", which shows the date the listing was last reviewed, ranges from 2012 through 2018. This means that some listings are no longer active, and a price from say 2012 is misleading as prices were cheaper at the time. Therefore we remove listings for which the last review date is prior to 2017.

We are interested only in the number of reviews for a property between 2017 and 2018. So, for 'reviews.csv', we grouped by 'listing id', applied aggregate function and merged them so that we get the total number of reviews for each listing id between 2017 to 2018. However, note that this newly calculated number of reviews column could still introduce bias to our analysis. If a property is listed many years ago whereas the other property is listed starting August 2018, the former property will have more reviews than the latter. To address this issue in the data, the following subsection describes our new column 'average number of reviews per month'.

3.1.4 Average Number of Reviews per month

To reduce the bias in our model, we introduce a new column called 'avg_reviews_per_month', which is calculated from the 'date' in 'reviews.csv' and the 'number_of_reviews' from 'listings.csv'. We took the first reviews as the month we considered the listings opened, and September 2018 as the last month as our data is was updated until that month. Then to calculate the new column, we find the number of months in between the two and divide the total number of reviews by it.

Obtaining this new column of data is significant because if a property is listed from 2017 January whereas the other property is listed from 2018 August, the former property will have more number of reviews than the later which makes our analysis biased. The new column represents the average number of reviews per month for each listing over the considered time period. We use this newly calculated average number of reviews per month in our analysis instead of 'number of reviews' column.

Finally, we added the 'average number of reviews per month' described above as a new column to the listings dataset by joining the two datasets with common column 'listing_id' followed by dropping the unnecessary columns. There are a few null values in 'avg_number of reviews' column because there are no reviews for those listings; we have filled these null values with 0. We call the processed

dataset 'updated_listings', which is used for fitting prediction models. Table 1 shows the columns in 'updated_listings' and the data type in each column.

Columns	Data type
listing_id	numerical
neighbourhood_group	categorical
room_type	categorical
price	numerical
average_reviews_per_month	numerical
availability_365	numerical

Table 1: Columns in updated_listings and data types

4 The Proposed Method

4.1 Assumptions

To predict future prices and popularities, we made the following assumptions:

- The macroeconomics conditions remain the same over the next few years, therefore Airbnb prices stay in the same range. Hence our prediction model is valid when used in the upcoming years.
- Number of reviews and number of days the property is booked for the next year is an indicator of the popularity to a location in listings.

4.2 Estimating Popularity of Neighborhoods

We evaluate the popularity of neighborhoods using the average number of days the listings in each neighborhood is booked over the next year. We calculate this value from availability_365 column which shows the number of days the listing is still available for booking next year. The calculation is simply 365 minus availability_365. We also use the 'average number of reviews per month' for listings in each neighborhood for popularity estimation. We then calculate average days booked in each neighborhood by taking the averages of all listings in each neighborhood. We calculated the neighborhood average number of listings per month in a similar fashion. The results are shown in the neighborhood popularity table in Fig.2 and Fig.3.

4.3 Price Prediction Model Development

We used regression techniques to build our learning model using the processed data. The main reason for choosing regression method is that the prediction target "price" is continuous valued. Given the nature of our target, binary or multi-class classification are not appropriate. In the model building stage, we

first transformed the 'price' column by taking natural log of 'price', which is the target for prediction. We then applied one-hot encoding to categorical variables, namely "neighborhood_group" and "room_type", to transform these columns to to numerical. The data is now ready for regression. We used Linear Regression, Lasso Regression, Ridge Regression, Support Vector Regression, and Random Forest Regression to build prediction models. The results are reported in Table 2. and Table 3 in the following section of this report.

To evaluate the performance of the regression models, we use Root Mean Squared Error (RMSE) as the validation metric.

5 Experiment

For our experiments, we use data obtained from the various processing steps described in Section 3 above. We build models on different regression algorithms like Linear Regression, Ridge, Lasso, Random Forest Regression and Support Vector Regression (SVR). To obtain better results, outliers are removed from the data. We consider the different standard deviations thresholds in removing outliers to see which gives the better results. We split the data as training set and test set with 80% of data as training data and 20% as test data.

5.1 Encoding Categorical Variables

Our data has two categorical variables 'neighbourhood_group' and 'room_type'. Since the machine learning algorithms work only on numerical data, we need to encode these variables. But, using just LabelEncoder is not enough for our model as the algorithm might consider that these variables have some order and gives one variable priority over other. As neighborhoods and room types categories do not have priorities, we use One-Hot Encoding, which generates one column for every category in each categorical variables.

5.2 Dummy Variable Trap

Once we encoded the categorical variables, we need to be aware of the dummy variable trap which could cause multicollinearity in our features. To eliminate this dummy variable trap, we eliminated one column for each categorical variable.

5.3 Log of Price

As the range of price variable is quite large, it is necessary to apply data transformation in order to obtain decent results. With such large range of values it is difficult to build a good model with any machine learning algorithm. Therefore we apply natural log to the 'price' values and we use log_price for our analysis and prediction.

5.4 Standard deviation

We also observe that there are outliers or rather very large values in each neighbourhood group. It is important to remove these outliers as including outliers in the model decreases the model performance. So, we consider to remove these outliers on the basis of standard deviation. We consider the values in the range of 0.5 standard deviation, 0.75 standard deviation, 1 standard deviation and 2 standard deviation from the mean. With 1 standard deviation, 68% of the data is retained and for 2 standard deviation 95% of the data is retained for building the model.

5.5 Variations of Standard deviation

We apply different regression algorithms to build our model and observe that Linear Regression produces decent results on 0.75 standard deviation. At 0.50 standard deviation we simply removed too much data despite producing the lowest RMSE. Hence we will not consider 0.50 standard deviation.

By taking this as a baseline, we apply Linear regression on different ranges of values of our data. We validated the performance of our models by calculating Root Mean Squared Error (RMSE) of every model. We tabulate the values of RMSE for all our models below in Tables 2 and 3.

Models	Standard Deviation	RMSE
Linear Regression	0.75	0.312
Ridge	0.75	0.314
Lasso	0.75	0.315
Random Forest	0.75	0.328
SVR	0.75	0.313

Table 2: Best Accuracy Obtained from Log of Price

Standard Deviations	RMSE
2	0.412
1	0.347
0.75	0.312
0.5	0.256

Table 3: Accuracy Obtained by Best Model (Linear Regression)

5.6 Results

5.6.1 Price Prediction

We decide to consider range of values with 0.75 standard deviation away from mean as we consider as not eliminating too much data and the RMSE value is

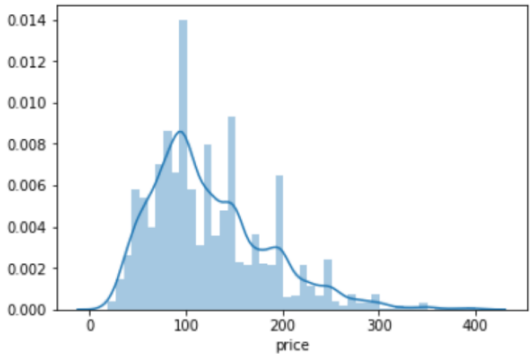

Figure 1: Distribution of Price of Data Within 0.75 Standard Deviation

also quite low. We also decide to go with Linear Regression considering it's low RMSE value compared to other regression algorithms.

Linear Regression Equation:

$$\text{log\_price} = \alpha + \beta_1 \times neighbourhood\_group + \beta_2 \times room\_type + \beta_3 \times availability\_365 + \beta_4 \times avg\_number\_of\_reviews$$

We run each model 10 times and considered the average values for better precision. This model gives an RMSE value of 0.312 which is quite good. The intercept and coefficients of the model are:

Intercept: 4.86849763e+12

Coefficients: [-5.77648706e+11 -5.77648706e+11 -5.77648706e+11
-5.77648706e+11 -5.77648706e+11 -5.77648706e+11
-5.77648706e+11 -5.77648706e+11 -5.77648706e+11
-5.77648706e+11 -5.77648706e+11 -5.77648706e+11
-5.77648706e+11 -5.77648706e+11 -5.77648706e+11
-5.77648706e+11 -5.77648706e+11 -4.29084892e+12
-4.29084892e+12 -4.29084892e+12 6.38227986e-05
-2.00402773e-02]

5.6.2 Popularity

We estimate popularity on the level of different neighbourhood groups. We define popularity based on the average number of reviews for different listings in a neighbourhood and average number of days the listings are booked for next year in the neighbourhood. We group our data on these two values and popularity ranking is given to different neighbourhoods in Seattle area. Fig 2 shows popularity of neighborhoods in decreasing order. In Fig 3, the neighborhood groups in the first quadrant are popular. The neighborhoods in the fourth quadrant are not popular. The remaining are somewhere in the middle.

neighbourhood_group	neighbourhood_avg_monthly_reviews	neighbourhood_days_booked_365
Interbay	3.4	288.2
Beacon Hill	3.2	237.7
Capitol Hill	3.2	221.5
Central Area	3.2	251.7
Queen Anne	3.1	228.0
Ballard	3.0	234.8
West Seattle	3.0	206.2
Delridge	3.0	221.7
Rainier Valley	2.9	221.8
Other neighborhoods	2.7	221.5
Downtown	2.5	200.7
Magnolia	2.5	201.6
Northgate	2.5	210.8
Lake City	2.4	229.7
Seward Park	2.3	202.3
Cascade	2.2	238.3
University District	2.0	178.2

Figure 2: Neighbourhoods popularity

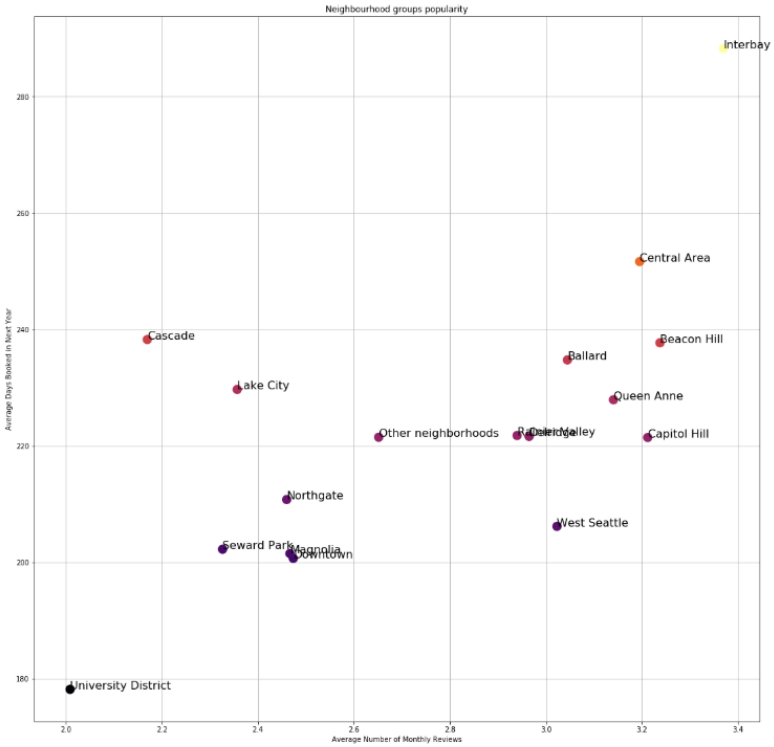

Figure 3: Popularity Among Neighbourhoods

5.6.3 Correlation & Linearity

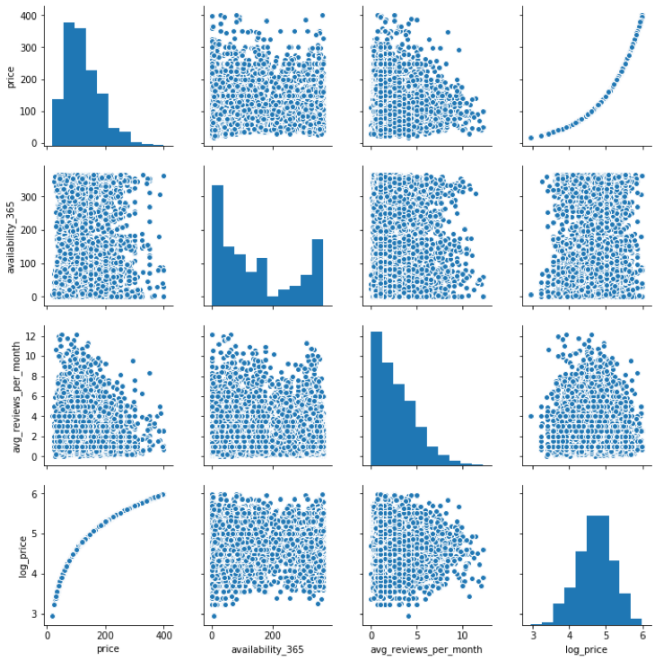

Figure 4: Relationship Among Features

As we see in Fig. 4, we cannot visualize a linear relationship among our independent and dependent variables. However considering the prediction model we obtain from Linear Regression with RMSE of 0.312, we believe that the linear relationship exists between the encoded categorical variables and price.

6 Conclusion

Through a series of data processing steps and model fitting, we concluded that linear regression is the best model for our data, which produced an RMSE of 0.312.

Our major challenge while building this model is the data. Different clusters of data are overlapped over one another hence, it is hard for the algorithms to differentiate and predict the data. Another issue was that we did not find significant linear relationship between any numerical independent variable and price. To overcome this challenge to some extent, we filtered out outliers through applying a standard deviations to the price column of the data and applied natural log transformation to the dependent variable. We also one-hot encode the categorical features. By applying all these transformations, we are able to achieve better results than what we would have obtained with the raw data.

References

[1] Inside Airbnb. Adding data to the debate. http://insideairbnb.com/get-the-data.html, 2018.

[2] Philip Mohun. Making models (i) airbnb price prediction: Data analysis. https://medium.com/@philmohun/making-models-airbnb-price-prediction-data-analysis-15b9af87c9d8, 2018.

[3] Nicole Samrao. Analysis of airbnb listings in u.s. cities - predicting listing price using the zillow home value index. github.com/nsamrao/Airbnb/blob/master/Airbnb%20Analysis%20Capstone%20Final%20Paper.pdf, 2017.