AirBnB Price Prediction Term Paper

Aniket K Singh

April 3, 2021

Airbnb Price Prediction

1. Introduction

Airbnb has become one of the essential elements of trips and vacation plans for over 150 million people. Since 2008, guests and hosts have used Airbnb for a unique and personalized experience of traveling with a wide range of travel possibilities. Before Airbnb, most consumers had to rely on hotels. As hotels aren't as widely available as Airbnb with their exceptional business model, they soon became the best vacation rental marketplace.
Because of the dramatic growth of Airbnb, price prediction becomes one of the essential elements for their platform. As hosts typically determine the price of the Airbnb. Both the host and Airbnb need to provide a fair price to the consumer, as it is an essential element of their model. Determining the price is crucial for the new and existing host/Airbnb because the price cannot be too high that they lose popularity or not get any guest. As for customers, they have options to check and compare prices depending on their needs.

2. Data Preparation

The dataset used in the project has been accessed from Kaggle's database. It is a public dataset of Airbnb that is accessible publicly on their original dataset website. This dataset describes the listing activity and metrics in NYC, NY for the year 2019. This data file includes all the needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions. The dataset is in Comma-Separated Values(.csv) file format. After importing the datset, we performed some analysis on the dataset. The dataset contained 48894 rows and 16 columns.

2.1. Handling Missing Values

After importaing the dataset. we noticed that column "last_review", "reviews_per_month"

had 10052 missing values. Also, few columns had some missing values that were either imputed or dropped. This dataset did not contain many missing values or the rows that contained majority of missing values wasn't used to train or build the model.

2.2. Dealing with categorical inputs:

Too many levels of a categorical variable are one of the most frequently occurring problems in predictive modeling. As for our dataset, multiple level of categorical variables. Three columns categories of the categorical variables are neighborhood_group, neighborhood and room_type are. In case of neighborhood_group and room_type, they do not have multiple levels so it is possible to use dummy coding. But in case of neighborhood, we have 219 levels which makes dummy coding overly complex. The number of dummy variables for the neighborhood variable would be one less than the number of levels in neighborhood. It would lead to the problem of high dimensionality which will ultimately cause the over-fitting of the data. Thus, the model will not be able to generalize properly to a new dataset. Moreover, inclusion of categorical variables with too many inputs can lead to the problem of quasi-complete separation. The problem of quasi-complete separation occurs when a level of the categorical variable has a target response of either 0% or 100% and can cause the interpretation of the regression model. However, we did not include few variables such as neighborhood in our final model.

2.3. Reducing redundancy:

#[this part is not done yet] Datasets with too many input variables can cause different problems in development of predictive models. For instance, the probability of redundant variables in a dataset with more input variables is high. In high dimensional datasets, the association between the inputs makes it difficult to identify the best predictor variables. So, we can reduce the redundancy by clustering numeric variables in the training dataset. Finally, we can use criteria such as 1-Rsquare statistic, relationship between predictor variable and the target, and subject matter knowledge to choose a variable from each cluster. We can use technique such as K-means clustering or KNN to create clusters for training dataset. #Please rember to do this before creating final model.

2.4. Variable screening:

In predictive modeling, carefully selected features can improve the accuary of the model while adding too many or too less variables can lead to overfitting or underfitting. The model can't generalize well and work on unseen dataset if the model is not "just right" or close to "just right". There are multiple methods for variable screening. We can perform many statistical tests such as AIC, BIC, F-test, Cross Validation to check what set of variables perform well on the model. There are many methods available such as Stepwise

selection which uses either Forward or Backward selection method. But, we have a very small number of features, so we can perform Best Subset selection method. Best subset selection method can be computationally expensive but in our case, it is possible to perform best subset.
#best subset yet to be performed

3. Models Used:

#I plan to introduce all the models that I would be using in the project.
#After Introducing model, I would make sections for Model Analysis and Performance

Linear Regression

Linear regression is a technique used to model the relationships between observed variables. The idea behind simple linear regression is to "fit" the observations of two variables into a linear relationship between them. Graphically, the task is to draw the line that is "best-fitting" or "closest" to the points (x_i,y_i), where x_i and y_i are observations of the two variables which are expected to depend linearly on each other.