The Model of Boston Airbnb Rental Prices Based on Stepwise Regression

Jinqian Pan | panjin@kean.edu | CPS3320

Background:

Boston is the capital and largest city of Massachusetts, and is one of the oldest and most culturally valuable cities in the United States. Whether it's a long stay or a short trip, Boston is an ideal place to go. After choosing to travel to Boston, how to choose the right local residence becomes a question worth considering. This project analyzes Boston rent data provided by Airbnb to build a model that targets rent prices.

Hypothesis:

The model of price using the Stepwise Regression will get about 90% accuracy.

Data:

The data set contains over 3 thousand airbnb listings in Boston from 2008 to November 2019. As part of the Airbnb Inside initiative, this dataset describes the listing activity of homestays in Boston, MA. This dataset is part of Airbnb Inside, and the original source can be found in http://insideairbnb.com/get-the-data.html.

The data set has 95 variables describing the information about more than 3,000 houses in all aspects, such as *bed_type*, *neighborhood*, *city* and *state*.

For some variable of data, the dataset shows “Not Applicable” (NA), which may increase the difficulty of building model. And the dataset also offers some data which is useless at analysis, such as *host_thumbnail_url* and *thumbnail_url*; these links could not be used when we build the model. So that the useful variables must be less than 95.

Analysis Steps:

Data Cleaning is the most import step for every analysis; we would reject useless data which mentioned above. And then we use Bar Plot to get the NA’s conditions of each column. After that, these NAs need to be handle, through deletes, decision trees, etc.

Secondly, using data visualization is a good way to get a number of informations from data which is given. For numbers of data, Box plot, Histogram, Stacked Chart and Line Chart are my choices to gain information. For string of data, we can use One Hot Encoding and Label Encoding to transfer string to number (using Dummy Variable is also a good way to transfer this data). After that, we can get the graphs from the string of data.

Finally, using Stepwise Regression to get the model of the Boston Airbnb rental prices.

Potential Roadblocks:

There are 95 variable which is too much to building the model, so that we need to give up some of them; however, giving up which variable is a big headache. And using Stepwise Regression also have other problem. Because of the rule of Multiple Regression, we need to figure out which variable could be Polynomial variable of the model.