The Model of Boston Airbnb Rental Prices Based on Stepwise Regression

Jinqian Pan | panjin@kean.edu | CPS3320
https://github.com/JinqianPan/CPS-3320/tree/master/Program01

Result of the project:

$$\begin{aligned} \text{price} = 104.6274 &- 15.4611 * \text{host\_identity\_verified} \\ &- 2.6295 * \text{property\_type} \\ &- 79.2308 * \text{room\_type} \\ &+ 10.3631 * \text{accommodates} \\ &+ 37.9868 * \text{bathrooms} \\ &+ 33.0472 * \text{bedrooms} \end{aligned}$$

where

host_identity_verified: whether host's identity is verified (0 means false, 1 means true)

property_type: the type of house (0-12 respectively means 'Apartment', 'Bed & Breakfast', 'Boat', 'Camper/RV', 'Condominium', 'Dorm', 'Entire Floor', 'Guesthouse', 'House', 'Loft', 'Other', 'Townhouse', 'Villa')

room_type: the type of room (0-2 respectively means 'Entire home/apt', 'Private room', 'Shared room')

accommodates: how many people can be contain in this house

bathrooms: the number of bathroom

bedrooms: the number of bedroom

It do not really follow proposal's "The model of price using the Stepwise Regression will get about 90% accuracy." But I already get the the model of Boston Airbnb Rental Prices. There are some reasons which can explain why I do not follow the proposal:

1. Time.
2. The data just have less than 4,000 records which means if I use 20% even 10% records as the testing set, my model cannot use such less training set to get the accuracy model.
3. In my opinion, the main point of this project is using the Stepwise Regression to get the model.

Analysis Step (different with proposal):

1. For data cleaning, it needed to use Bar Plot to get the NA's conditions of each column; however, I find a new easy way which can replace it (just use isnull().sum() method). This method can also show the number of the NA with numbers.

2. For data visualization part, after knowing the information of data, I just draw a small number of graph for this data. Because of the properties of this set of data, it does not have great value to get the graph for the data.
3. In the project, I just use Label Encoding to transfer string to number, because there are too much string variable in my model; if I use the One Hot Encoding, the model might be too long to see.

Challenge:

The challenge part is defining the function for stepwise regression in this project. In the R language which I use in my Data Analysis course, I can use packages and functions to do a variety of regressions. In most time, I just need one line code to analysis; however, I do not find packages which similar as R language in Python. It means I need to go to the website to find and understand someone's code or write by myself.

Reflection:

Splitting the data into testing set and training set would be the first step after data cleaning, if do this project again. Testing set can be used to test the model and get accuracy of the model. Just using accuracy as standard is also not the best way to make sure whether the model is fitting; next time I might use one of other methods (Accuracy, Precision, Recall or F1) as standard. And I will try to use polynomial regression to replace just single power variables.

In my opinion, I think I need to focus on coding than data analysis for this project, because python do not have lots of statistic package, so that I need to code by myself.