Predicting Airbnb Prices in Florence, Italy

Tanvi Hindwan,

(Group 5)

DATS6103: Data Mining

The George Washington University

December 3, 2019

Introduction

Airbnb, a company designed to let average people rent their homes and/or properties, is one of the largest in the world with over 7 million accommodations listed and over half a billion-guest check-ins (CITE - Airbnb). **What features lead to higher Airbnb rental prices?** To investigate this question, this research focuses on the Airbnb market in Florence, Italy.

Methods

Data Compilation and Data Importing:

The dataset used for this research was comprised of information from 2019 that described Airbnb rentals in Florence, Italy. The Florence, Italy dataset was chosen because it presented an opportunity to add to the existing analysis of Airbnb price predictors but for an under-analyzed location. The dataset was located on the Inside Airbnb website which contains a variety of Airbnb datasets that are available for public use through Creative Commons (CITE - Inside Airbnb). The raw Florence Airbnb dataset was 58MB and contained 105 variables for each for each of 12304 observations of rentals.

Data Modeling: Linear Regression

To check the linear relationship between target “price” and impacting features, data modeling using linear regression was performed. Linear Regression of data modeling to discover the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables) by splitting the whole dataset into training and testing sets and then fits the linear model using the scikit-learn library.

I have created 2 models, where I’m considering the best features using correlation matrix and all the features from the random forest for the linear regression.

Model1: Including the best features from the random forest using a correlation matrix into a dataset for regression

Firstly, correlation matrix was plotted to analyze which features are negatively and positively correlated coefficients with our target variable price.

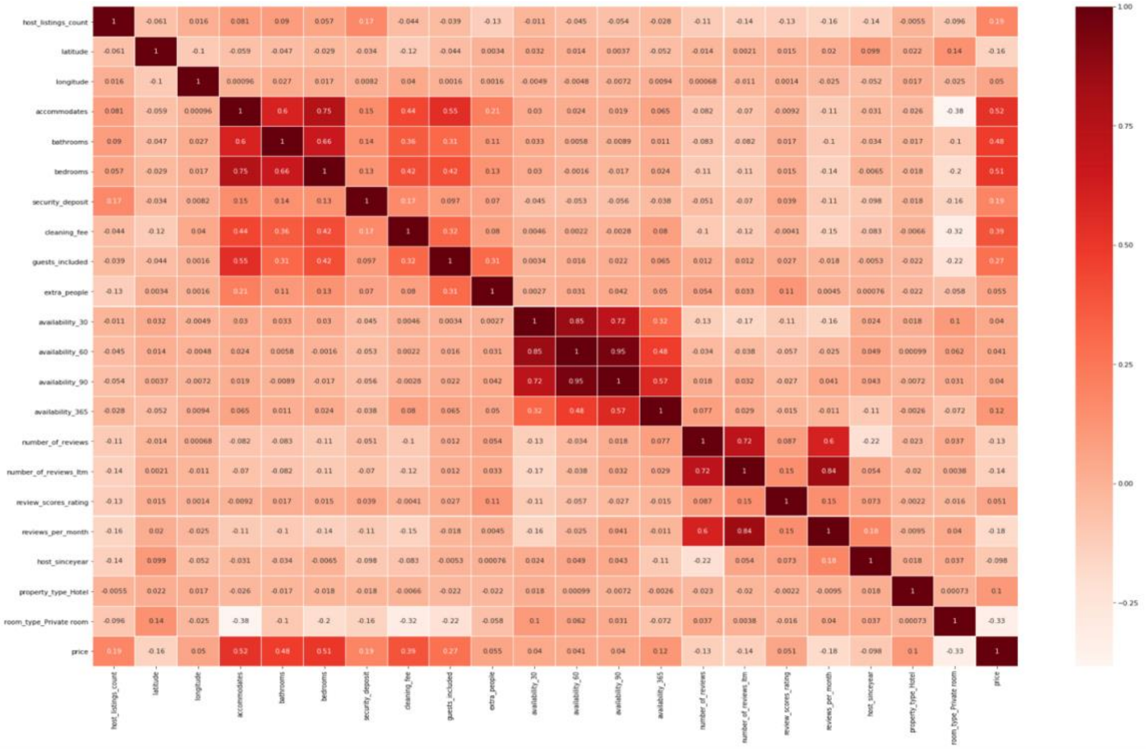


Figure 1. Feature Selection using Correlation plot

As can be seen from Figure 1, features like accommodates, bathrooms, bedrooms, security deposit, cleaning fee etc. are positively correlated with our target variable. So, I applied those positively correlated features to our linear regression model for checking the model's performance. For model evaluation, I split our dataset into training and testing set using function "train_test_split" from the scikit-learn library. Then, standardized our training and testing set to bring all the variables on the same scale. Once it was done, the linear regression model is applied upon the training set and predicted on the testing set to calculate the model’s performance on unseen data.

Model 2: Linear regression of all the features selected from the random forest feature selection.

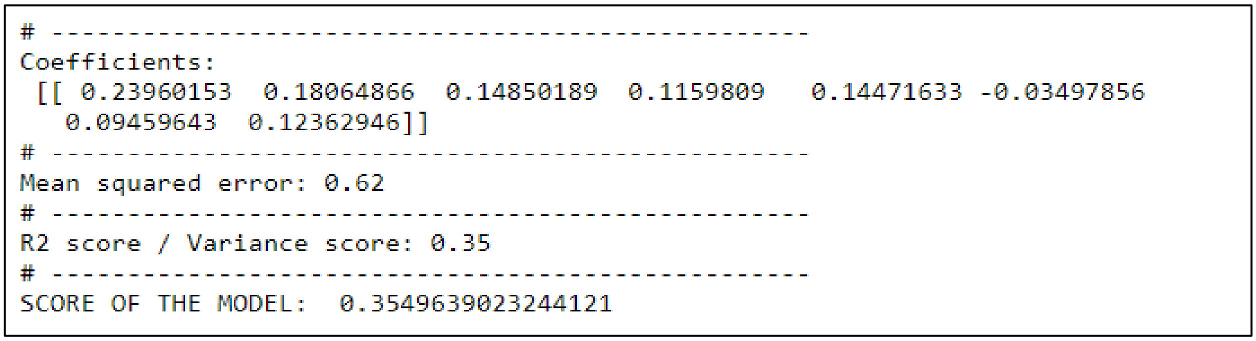
In this step, I took the whole dataset without finding the correlation with the target variable price and splitting the whole dataset into training and testing sets and then fit the linear model using scikit-learn library. The linear regression model was applied upon the training set and predicted on the testing set to calculate the model's performance on the unseen data like I have done in the previous model.

Results

I will put forward the results obtained from the research methods

Model Evaluation: Linear Regression Model

From the Model 1 linear regression results, I have considered positively correlated variables out of the correlation matrix. For the model evaluation, I have split our dataset into training and testing set using function “train_test_split” from the scikit-learn library. Then, I have standardized our training and testing set to bring all the variables on the same scale. I have used the scikit-learn to import the linear regression model. I have fitted the model on the training data and predicted the values for the testing data.



```
# --------------------------------------------------
Coefficients:
 [[ 0.23960153  0.18064866  0.14850189  0.1159809   0.14471633 -0.03497856
   0.09459643  0.12362946]]
# --------------------------------------------------
Mean squared error: 0.62
# --------------------------------------------------
R2 score / Variance score: 0.35
# --------------------------------------------------
SCORE OF THE MODEL:  0.3549639023244121
```

Figure 2. Model 1’s Performance

Figure 2 shows the accuracy of our model that is R-square came out as 35% and mean square value (cost function) came out to be 62% on the testing set. As discussed earlier in the report the linear regression fit the model on a training set and predict on the testing dataset. So, I have tried to visualize the differences between actual prices and predicted values.

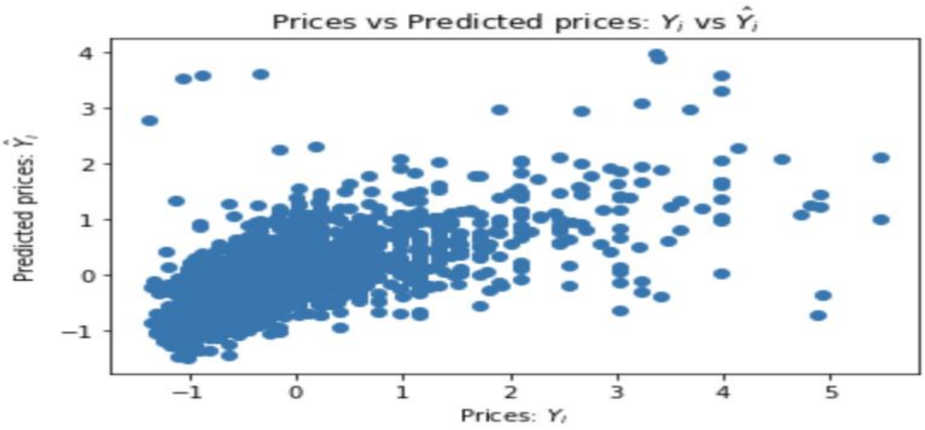
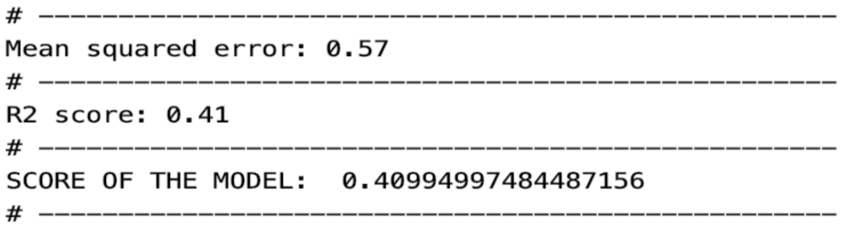


Figure 3. Scatter Plot of Model 1’s linear regression

Figure 3 shows the results that the model does not fit 100 %, as the scatter plot should create a linear line. These values are still acceptable because our kind of dataset has user described data and interactions.

From the Model 2 results, I have considered all the features selected from the random forest feature selection for the linear regression.



```
# ---------------------------------------------------
Mean squared error: 0.57
# ---------------------------------------------------
R2 score: 0.41
# ---------------------------------------------------
SCORE OF THE MODEL:  0.40994997484487156
# ---------------------------------------------------
```

Figure 4. Model 2's Performance

Figure 4 shows the accuracy of our model that is R-square came out as 41% and mean square value (cost function) came out to be 57% on the testing set Here, I have tried to visualize the differences between actual prices and predicted values for this model.

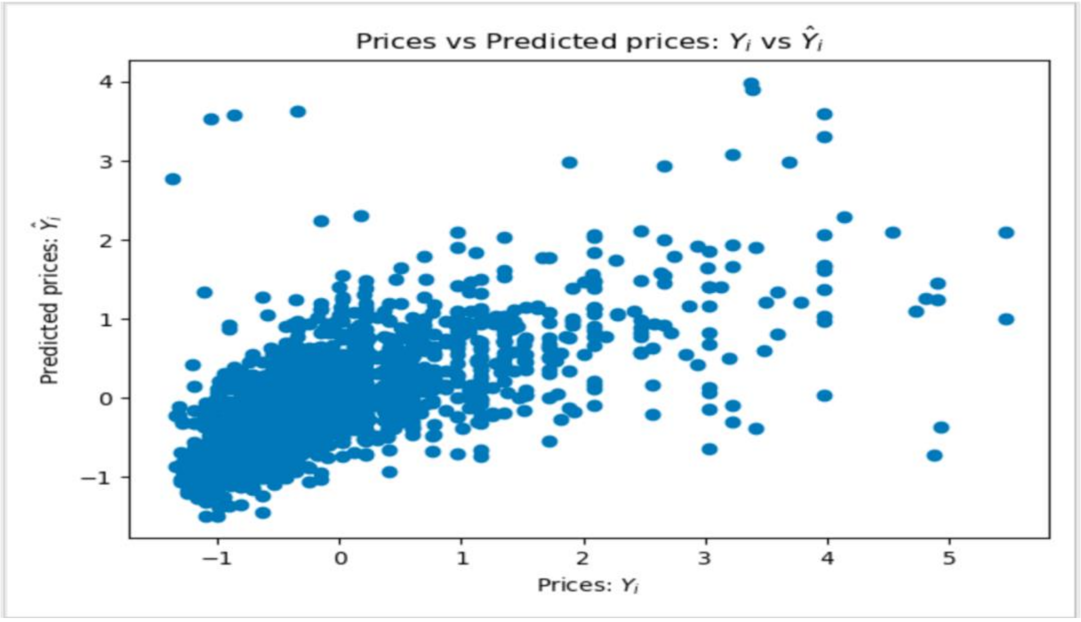


Figure 5. Scatter Plot of Model 2's linear regression

Figure 5 shows the results that the model does not fit 100 %, as we didn’t get the linear line. These values are still acceptable because our kind of dataset is socially biased.

Discussion

The purpose of this study was to determine the important features that have an impact on the prices of Airbnb. From the Model 1 data, I have found a few positively and wisely correlated variables to our target variable price. From the Model 2, I have found the data where the performance matrix has shown a similar value to Model 1. When the model is a good fit then the scatter plot should create a linear line. However, neither of the models has shown a good fit because the scatter plot is not creating a linear line in either of them. The R-square value of the Model 1 linear regression has come up to 35 %. and this value can be considered good for our dataset because our dataset is socially biased. Features like accommodates, bathrooms, bedrooms, security deposit, cleaning fee, guests included, 365 days availability and hotel property does contribute in making price prediction in our dataset.

For future work, I would like to include the following things:

a) A wider geographic area e.g. the rest of Italy or other big cities around the world.

b) Use a better-quality dataset with more accurate values for the listings. I will try to do data modeling with other Regularization techniques. Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting, which may result from simple linear regression.

c) The next big step is to augment the model with natural processing or machine learning.

References

1. Ismail, H. (2017) Learning Data Science: Day 9 - Linear Regression on Boston Housing Dataset
   Retrieved from https://medium.com/@haydar_ai/learning-data-science-day-9-linearregression-on-boston-housing-dataset-cd62a80775ef

2. Wikipedia. (n.d.) Linear Regression. Retrieved form
   https://en.wikipedia.org/wiki/Linear_regression