Predicting Airbnb Prices in Florence, Italy

Jyoti Sharma

(Group 5)

DATS6103: Data Mining

The George Washington University

December 3, 2019

Table of Contents

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 (Airbnb Inc., n.d). Renters vary from those looking to earn a little extra cash from a spare room to those who develop unique properties specifically for the Airbnb market. The type of rental, the amenities available, location, and other rental features may have an impact on the price a renter is able to charge while ensuring full occupancy. What features lead to higher Airbnb rental prices? To investigate this question, I contributed my part of work in data preprocessing and features selection for our research.

Methods

This section puts forward the research methods that I applied to understand the features impacting an unexplored Airbnb dataset of Florence, Italy.

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 raw Florence Airbnb dataset was 58MB and contained 105 variables for each for each of 12304 observations of rentals. Variables included aspects like the URL for each listing, amenities included, zip code, review scores, and number of beds/bathrooms, among many others. After importing the dataset as data frame using Pandas, it was discovered that the dataset contained various full text attributes that required natural language processing (NLP). Since NLP was not being used for this analysis, variables with full text attributes were removed. Table 1 shows the wider picture of features and the target that was the focus for further analysis.

Table 1
Missing values and Data Type of Variables in Dataset

Target	NaNs	Dtype	Features	NaNs	DType
price	0	object	host_since	5	datetime64[ns]
			host_response_time	1566	object
			host_response_rate	1566	object
			host_acceptance_rate	12304	float64
			host_is_superhost	5	object
			host_listings_count	5	float64
			neighbourhood	0	object
			latitude	0	float64
			longitude	0	float64
			property_type	0	object
			room_type	0	object
			accommodates	0	int64
			bathrooms	5	float64
			bedrooms	7	float64
			beds	23	float64
			bed_type	0	object
			amenities	0	object
			square_feet	11940	float64
			security_deposit	3010	object
			cleaning_fee	1766	object

guests_included	0	int64
extra_people	0	object
minimum_nights	0	int64
availability_30	0	int64
availability_60	0	int64
availability_90	0	int64
availability_365	0	int64
number_of_reviews	0	int64
number_of_reviews_ltm	0	int64
first_review	1663	datetime64[ns]
last_review	1663	datetime64[ns]
review_scores_rating	1736	float64
instant_bookable	0	object
cancellation_policy	0	object
reviews_per_month	1663	float64

Table 1

Data Cleaning

The dataset was first checked for whether there were missing values or outliers. The data types of variable were defined by category or numeric as per the requirement of analysis. For chosen dataset, I did various data pre-processing steps like data cleaning, data transformation, data manipulation, and data imputation.

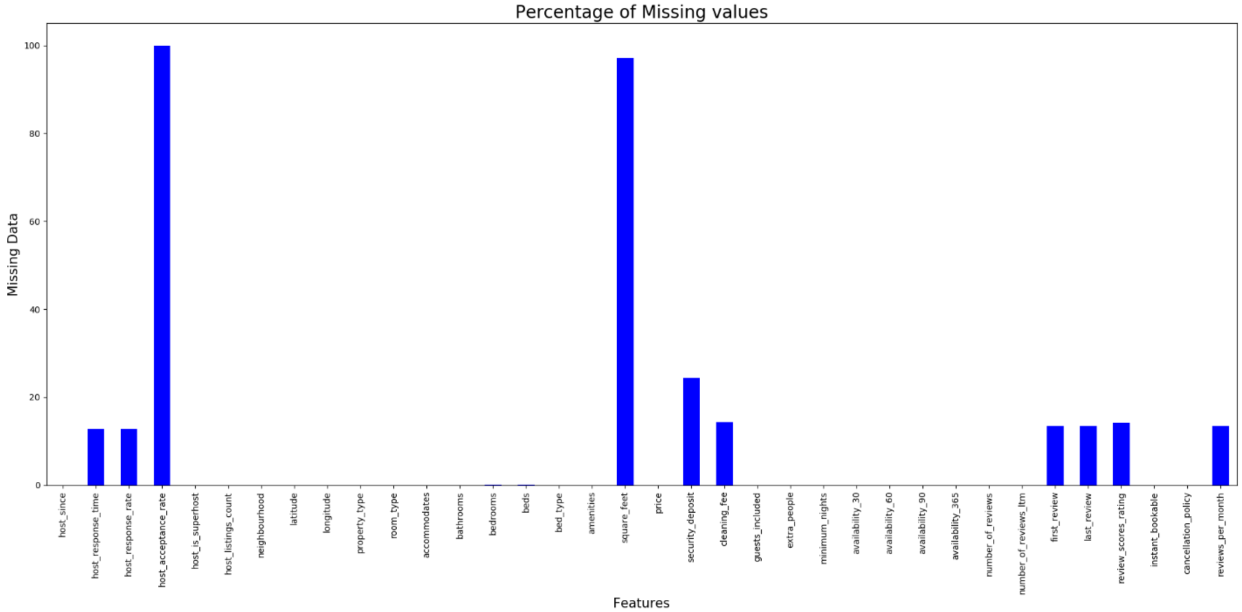

Figure 1. Graph for visualizing missing values in dataset

As, shown in Figure 1, there were many features with more than 30 - 40 percent missing values. Therefore, they were deleted using the “pd.drop” command from the pandas library After exploring the overall data, the target variable “price” was manipulated and cleaned using “.str.replace() and pd.to_numeric” commands to make it numeric dtype. Figure 2 shows the box plot for target “price”. This boxplot allowed for an investigation into outliers, as they could impact the accuracy of predictions.

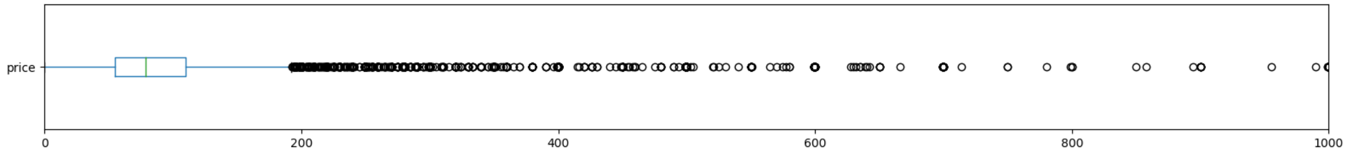

Figure 2. Outliers in target variable “price”

Among the other features, “amenities” were extracted and cleaned to see if it had any impact on price. Figure 3 shows top 10 amenities that were extracted for analysis.

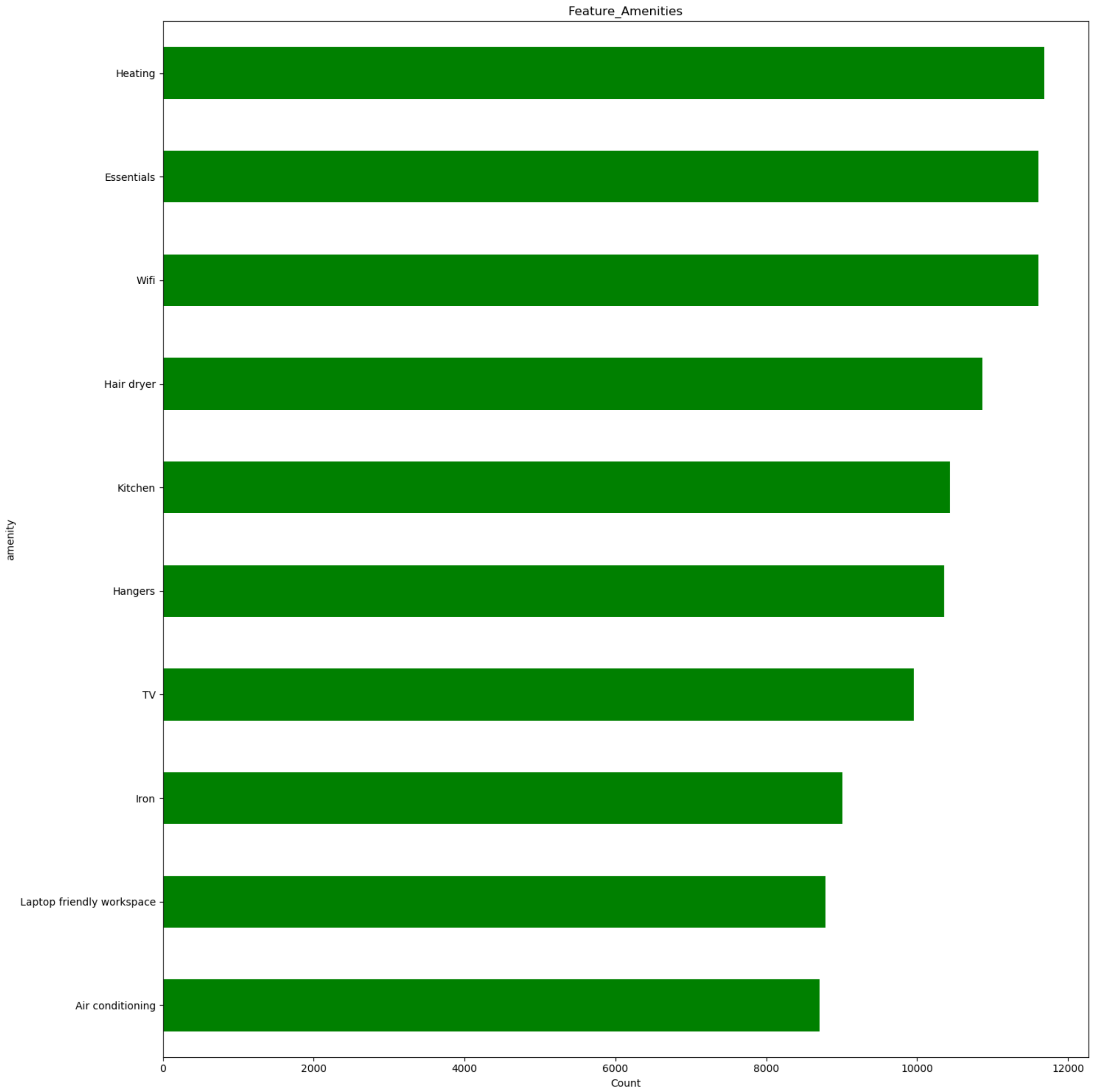

Figure 3. Top 10 amenities

To deal with the missing values median and mean imputation techniques were applied using “.fillna().mean() or. median()” from panda libraries. Finally, I applied “.dropna()” to deal with overall missing values. There were many categorical variables also that I also handled using one-hot encoding technique “pd.get_dummies” to create dummy variables.

Feature Selection:

Feature selection plays an important role for every data set that has hundreds of columns (variables). It is also known as dimensionality reduction. Among the various techniques, I applied Feature Importance technique using “RandomForestRegressor” and cross validated the same using another package called “SelectFromModel” from “sklearn” library. Therefore, I first divided whole dataset in to training and testing set and then fit the Random Forest model using “sklearn library” to extract the important features out of training set. Figure 4 shows list of important features that were found after applying RandomForestRegressor.

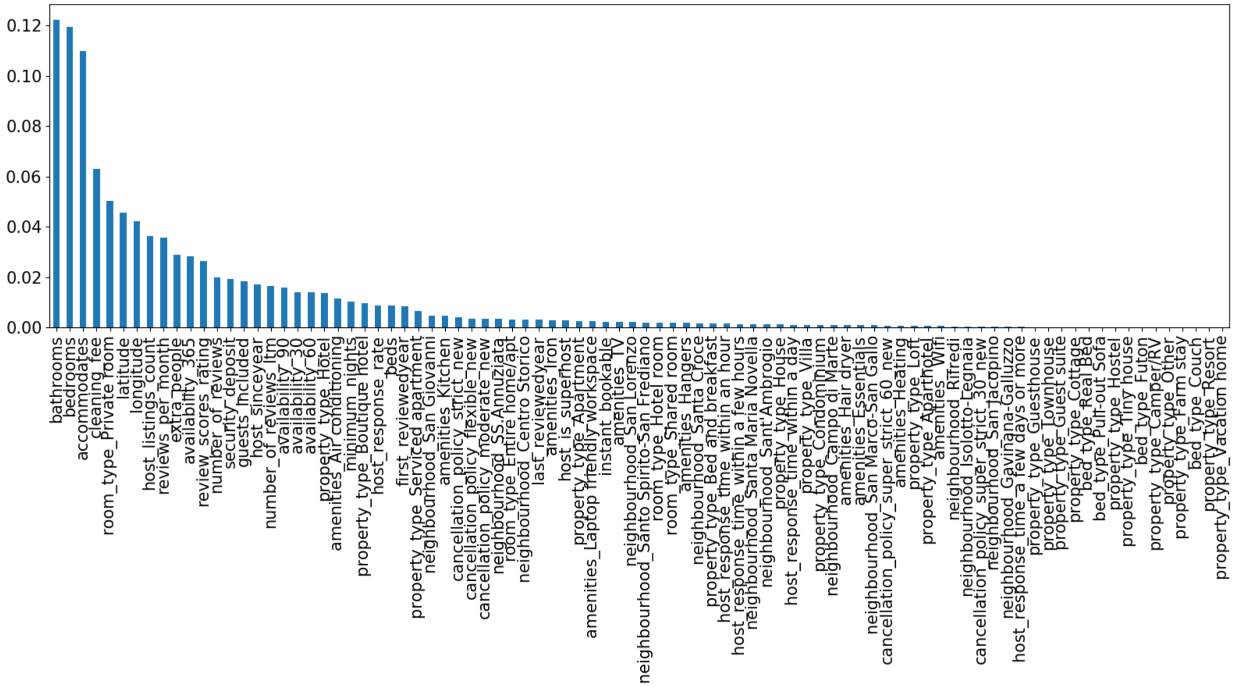

Figure 4. Important Features from RandomForestRegressor

Results

In this section, the results that were obtained from the research methods are examined for Florence, Italy Airbnb dataset.

Data Preprocessing:

In data preprocessing steps, when dealing with the target column “price”, I found some outliers in box plot. As, we know outliers are the extreme values in the set of data which can affect the mean value of the data thus I removed those using “pd.drop()” by setting the threshold value. Object type variables like amenities is extracted and transformed to series of dummy variables because we cannot deal with categorical variables to predict price directly. After executing all the preprocessing steps to make dataset ready for modeling, I tried to visualize correlation among all the features by plotting correlation plot (Figure 5) using heatmap (seaborn library). However, it was hard to check which features are important, and which were not due to so many features on correlation plot.

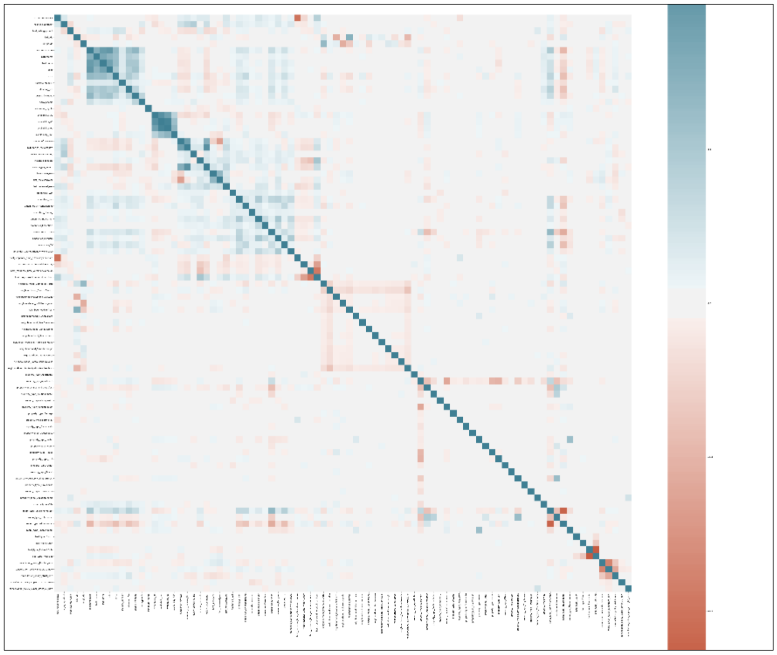

Figure 5. Correlation Score Plot for features

Dimensionality Reduction:

Random Forest Regressor is an ensemble technique which works under the hood with multiple regression trees outputting mean prediction of individual tree. Figure 6 shows the top 22 features that were found.

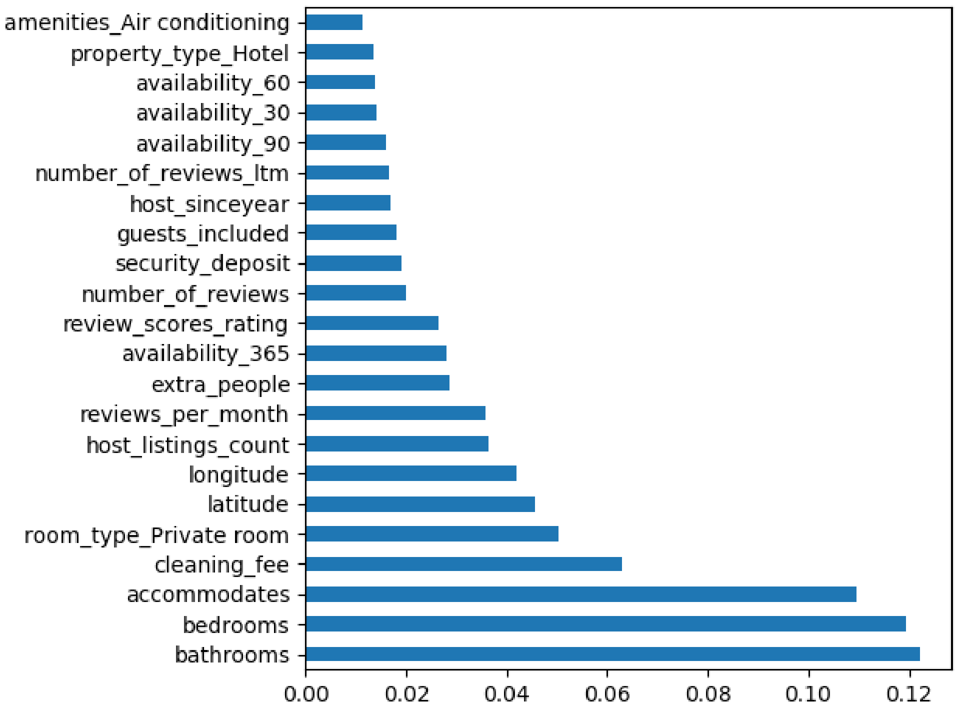

Figure 6. Top 22 features

I used two techniques to get the final important features. First, I applied RandomForestRegressor as ensemble technique from “sklearn” library where I performed training and got “feature_importances_” and in the order of descendance I estimated my top features. Secondly, I applied “SelectFromModel” meta transformer along with random forest. This technique automatically selected the good features from my training set. The beauty about this technique is that it outputs the array of boolean values (true for the features whose importance is greater than the mean importance and false for the rest). From the boolean value generated, I created list and got count and names of important features. So, this way I estimated my top 22 features (Figure 6). These features were then applied to make predictions for data.

Discussion

In this section, research is concluded, and key results are provided.

The purpose of this study was to determine the important features that have impact on prices of Airbnb's. Through, our methodology of feature selection we got around 22 features as important features however, very few features were found positively and wisely correlated with our target variable "price". Another, feature selction method called PCA (principle component analysis) was also tried however, it did not work for our dataset because probably the relationship between most of the variables were very weak (below 0.3). For future work, it would be interesting to assess the text attributes (which were avoided this time) to do text mining. The most frequently used word can be analyzed to know its importance in price predictions. As, the dataset is fully unexplored so many achievements can be accomplished to add the contribution to it.

Percentage of code refereed

As per the formula, the percentage of the code that I referred is 10%

References

Airbnb, Inc. (n.d.). Airbnb Newsroom: About Us. Retrieved November 25, 2019, from https://news.airbnb.com/en-in/about-us/

Inside Airbnb (n.d.). Get the Data. Retrieved November 25, 2019, from http://insideairbnb.com/get-the-data.html

Madden, G., Vicente, M.R., Rappoport, P., Banerjee, A., (2017). A contribution on the nature and treatment of missing data in large market surveys. Applied Economics, 49(22), 2179-2187. https://doi.org/10.1080/00036846.2016.1234699

Gupta, S. (2019). Airbnb Rental Listings Dataset Mining Retrieved from https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec

Akash., (2018). Feature Selection Using Random forest, The Wisdom of Crowds. *Towards Data Science* https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f