Predicting Airbnb Prices in Florence, Italy

Jyoti Sharma, Tanvi Hindwan, and Jessa Henderson
(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, this research focuses on the Airbnb market in Florence, Italy.

Methods

This section puts forward the research methods that have been 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. A variety of Airbnb datasets have been analyzed and shared online on sites, like Kaggle.com, for other locations, like Seattle and New York City. 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 (Inside Airbnb, n.d.). 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. According to Madden et al. (2017), many “nonresponse (or missing data) is often encountered in large-scale surveys” (p. 2179). Therefore, the data types were defined by category or numeric as per the requirement of analysis. For the Florence Airbnb dataset various data pre-processing steps were completed like data cleaning, data transformation, data manipulation, and data imputation.

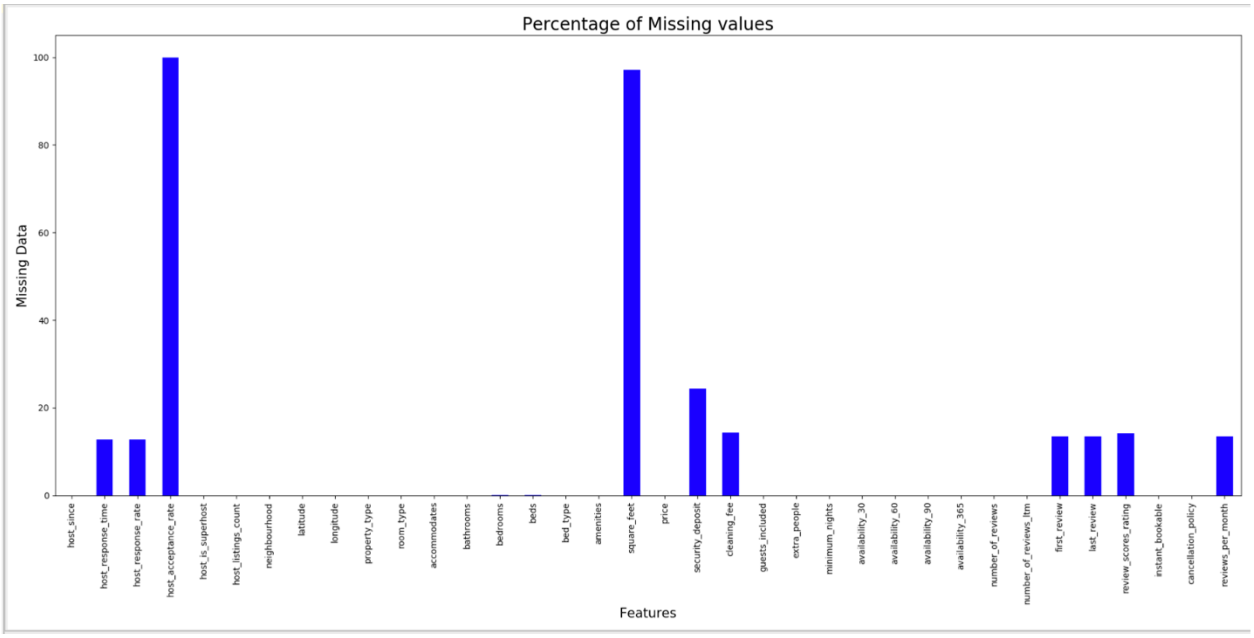


Figure 1. Graph for visualizing missing values in dataset

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.

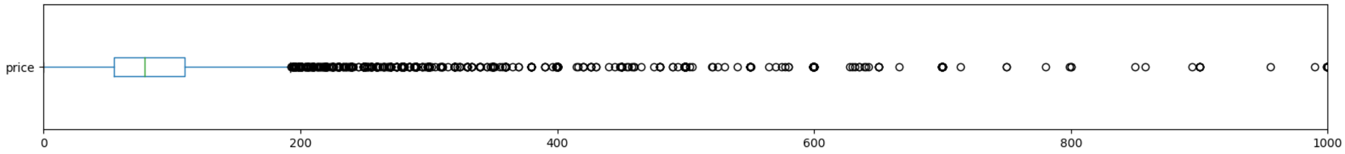


Figure 2. Outliers in target variable “price”

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

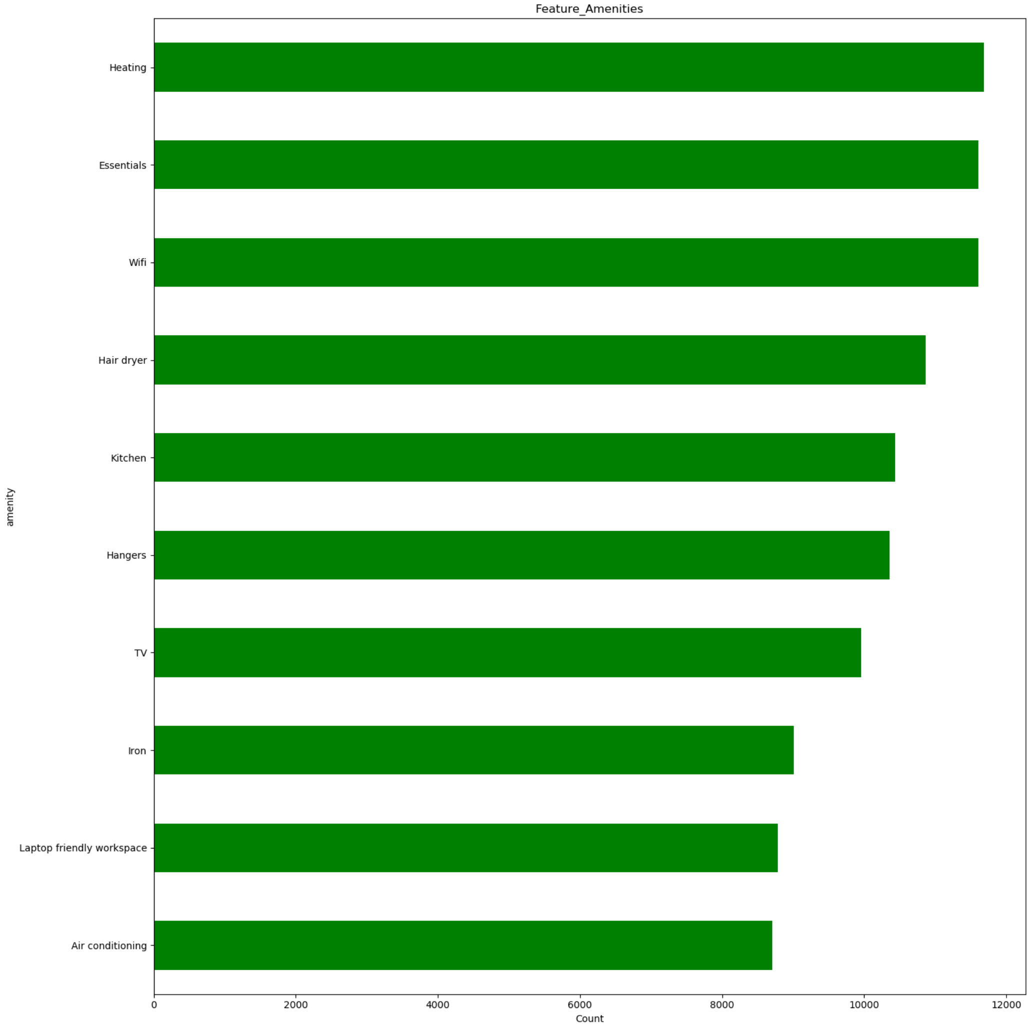


Figure 3. Top 10 amenities

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

Feature Selection:

The higher the number of features the harder it is to visualize the training set and to work upon. Therefore, feature selection plays an important role for every data set that has hundreds of columns (variables). It is also known as dimensionality reduction because it reduces the dimensions of our feature set. Among the various techniques, the Feature Importance technique was applied using “RandomForestRegressor” for feature_importances. The whole dataset was divided in to training and testing sets and then fit the model RandomForest using “sklearn library” to extract the important features out of training set. Figure 4 shows the list of important features that were found using “RandomForestRegressor” .

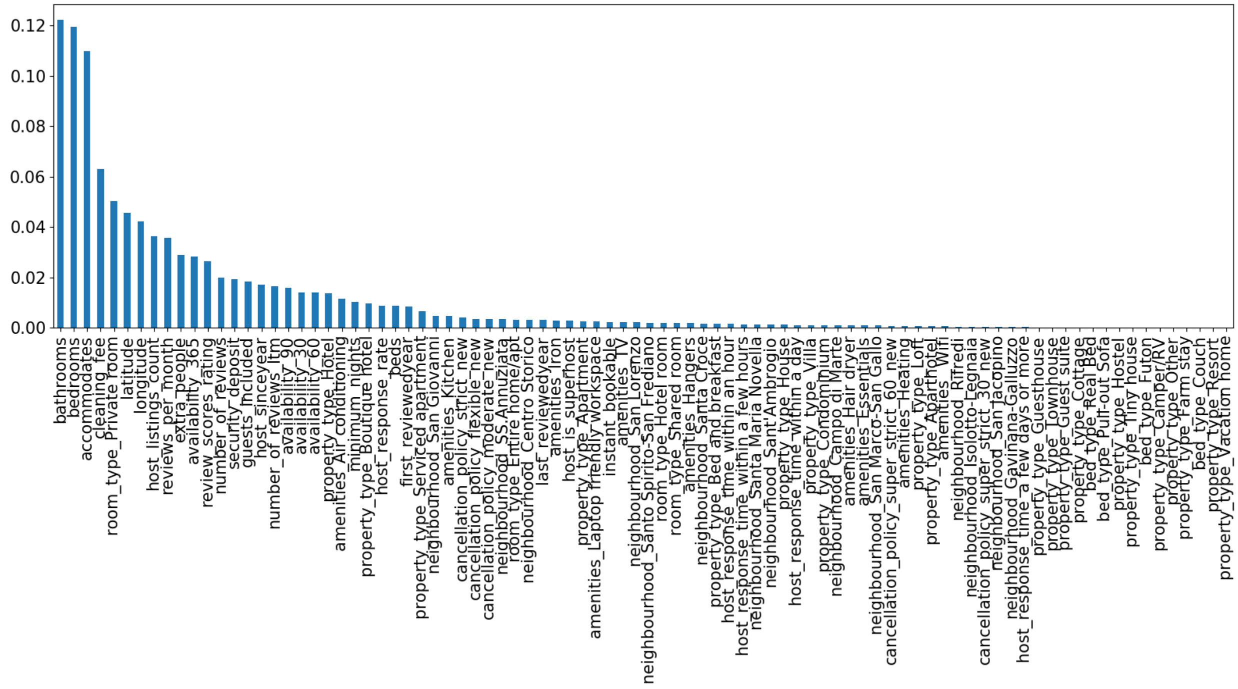


Figure 4. Feature Importance Graph

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 scikit-learn library.

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

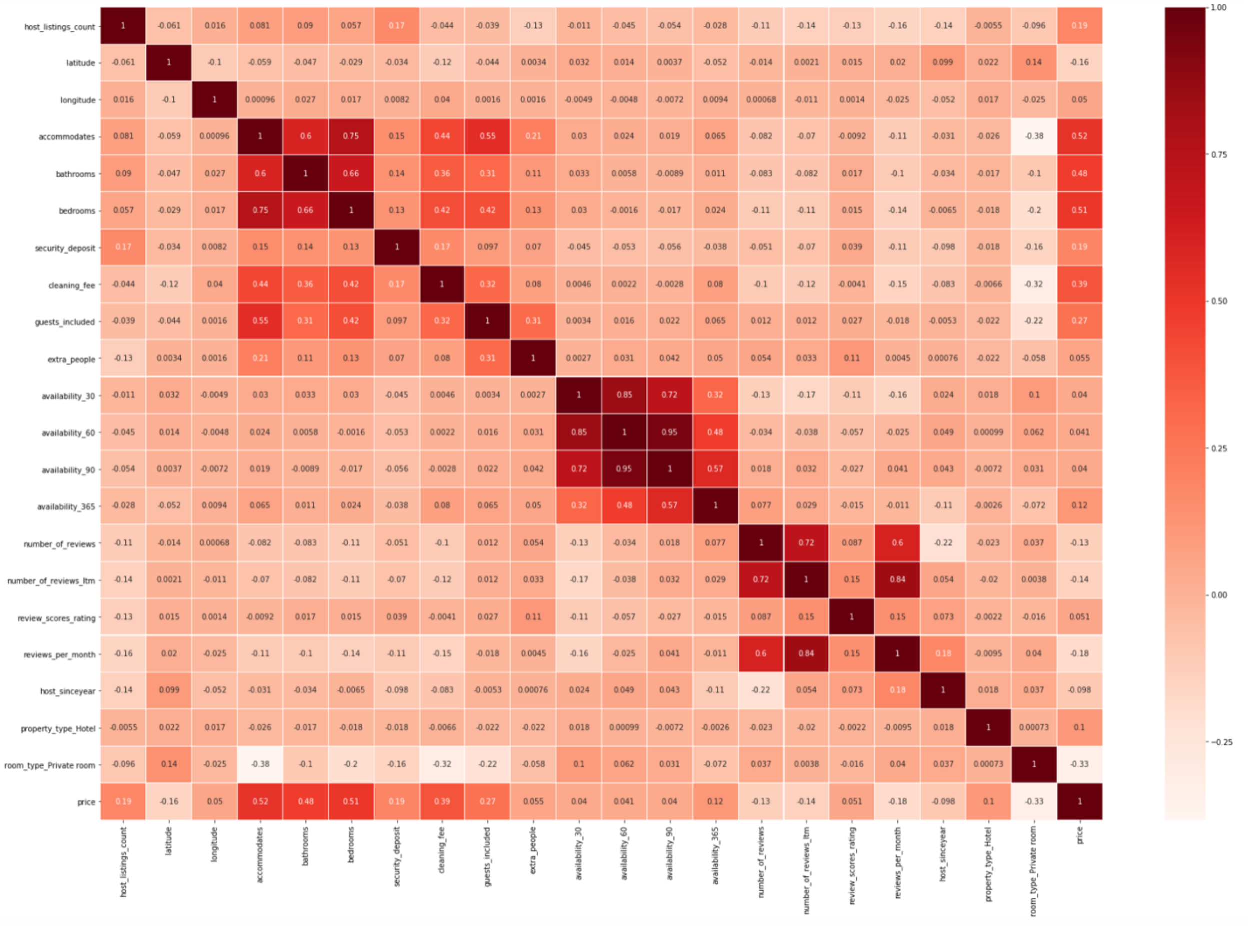


Figure 5. Feature Selection using Correlation plot

As seen in Figure 5 that only some of the features like accommodates', 'bathrooms', 'bedrooms', 'security deposit', 'cleaning fee', 'guests included', '365 days availability, 'Hotel property' were found positively correlated to our target variable. These positively correlated features were further used for modeling.

For model evaluation, we split our dataset into training and testing set using function “train_test_split” from scikit-learn library. Standardization using “StandardScalar” was also performed on to training and testing set to bring all the variable on the same scale. Linear regression model was applied upon training set and predicted on testing set to calculate model’s performance on unseen data.

Results

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

Random Forest Dimensionality Reduction:

Feature importance using “RandomForestRegressor” helped to extract important features out of the large training set. This way we got important features that could be applied for modeling. Figure 6 shows the top 22 features that were found as important. Cross validation of important features was also implemented using using “SelectFromModel” package. “SelectFromModel” is a meta-transformer which selects the important features based on their importance weights.

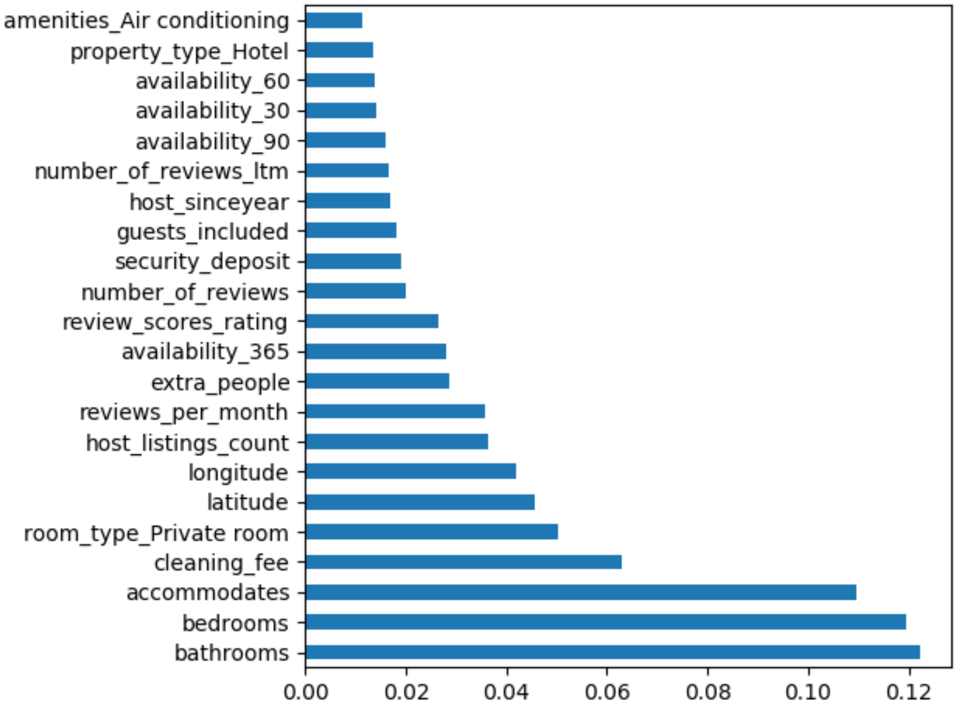
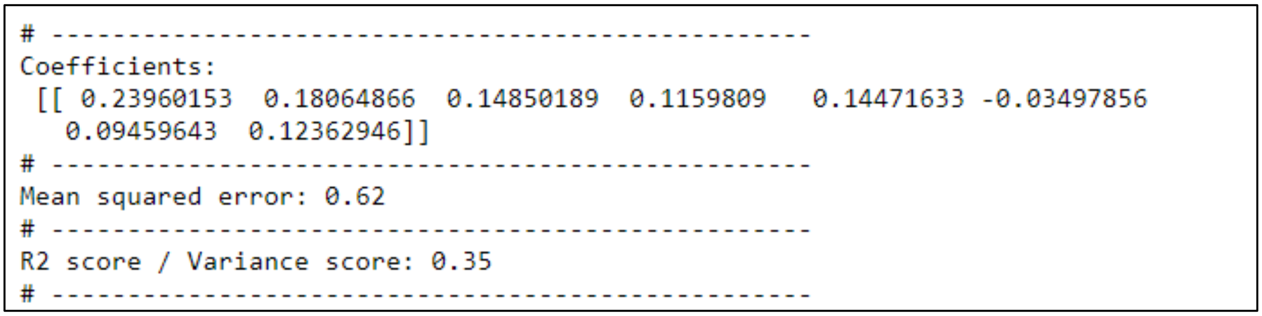


Figure 6. Top 22 features

Finally, Linear Regression modeling was conducted and evaluated for making model comparisons.

Model Evaluation: Linear Regression Model

During linear regression modeling, positively correlated variables were considered out of correlation matrix. For model evaluation, we fitted the model on training data and predicted the model's performance on 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
# ----------------------------------------------
```

Figure 7. Model's Performance

It can be seen in Figure 7 that the coefficient of determination (R-squared) was estimated to be 35 percent, which means the selected variable explain 35 percent of the variance in prices. Mean square error was estimated 62 on testing set, which is not bad. Scatter plot for better visualization of prediction on price was also plotted to check the how much actual price is differed from predicted values.

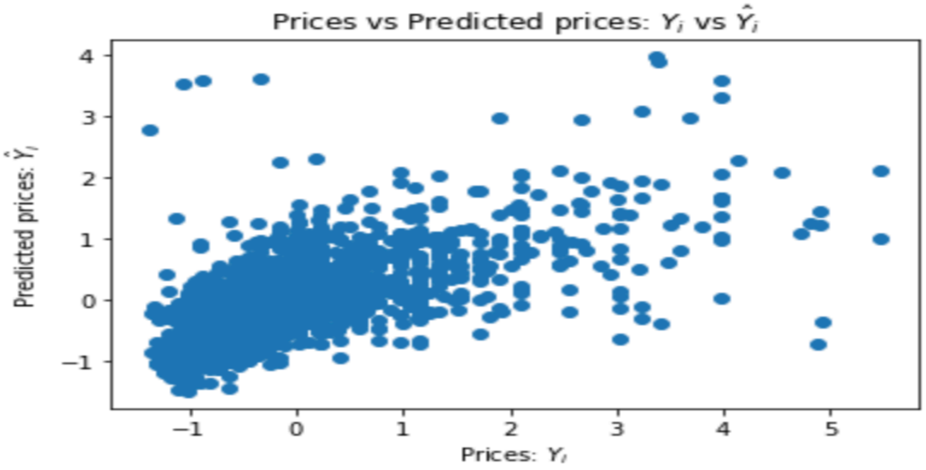


Figure 8. Scatter Plot of Linear Regression

It can be seen in Figure 8 that the model does not fit 100% as we didn't get linear line. However, for our kind of dataset where we have all user described data and interactions these values are still acceptable.

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 an impact on the prices of Airbnb's. The methodology of feature selection computed around 22 features as important features. Those features were cross verified using two different packages from "sklearn" library. However, very few features were found to be positively and wisely correlated with the target variable of “price”. Performance metrics evaluated after basic linear regression gave 35 percent coefficient of determination (R-squared) with mean square errors up to 62, which is not so bad for our kind of dataset. Scatter plot plotted to check the variance also showed the model does not fit 100%; however, this works for the kind of dataset worked with here. While concluding, itd can summarized that features like ‘accommodates’, ‘bathrooms’, ‘bedrooms’, ‘security deposit’, ‘cleaning fee’, ‘guests included’, ‘365 days availability’, ‘Hotel property’ were found to be contributing in making price prediction.

For future work we would want to, (a) include a wider geographic area e.g. the rest of Italy or other big cities around the world and (b) use a better-quality dataset with more accurate values for the listings. The next big step is to augment the model with natural processing or machine learning. With NLP we will run sentiment analysis for the listing’s reviews. We can also work on the model predicting the price for Airbnb with only the listings to help the owners to set prices for new locations removing features that are time restricted. Also, a content-based recommender system can be made for the customer to find the best Airbnb according to the prices.

References

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

Dubey, A. (2018). Feature Selection Using Random Forest. Retrieved from https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f

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

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

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