**DSCI-633 Foundations of Data Science and Analytics**

**Final Project Report**

# PREDICTION OF PRICES IN AIRBNB LISTINGS


**Project Members:**

Nisarga Khairnar -nk1976@rit.edu

Praharshita Kaithepalli-pk2971@rit.edu

**Course Instructor**:

Dr.Nidhi Rastogi

**Index**

## 1.Introduction:

In this project we have decided to use the AirBNB data set to predict the prices of different AirBNB listings based upon various other features available in the data set. The data was taken from a website named www.insideairbnb.com . This is because the AirBNB company will never release data from their side. In this particular data set all of the information was taken from the AirBNB website that is viewable by the users and put together in the form of a csv format. We have decided to go forward with the New York city AirBNB dataset as it is a very popular tourist destination and a business hub with a lot of AirBNB listings that are majorly booked throughout the year. This data set included listings information that was last scraped in November 2020, so this is a relatively new dataset.

## 2.Development of question/Hypothesis:

We know that the price of any hotel/AirBNB listing will decide upon various factors and it is not a random value set up by the host/owners. We have information about the location, host and review scores and the description of the listing(like number of bedrooms, bathrooms etc). So using these features we will attempt to predict the price based upon these features available to us.

## 3.Literature Review:

For visualization we use various scatter plots, bar plots, histograms and correlation heatmaps to find the correlation between different features. In this project we shall use various regression models and one KNN model to try and predict the price of the AirBNB listing. Some of the regression models we used are Linear Regressor, Random Forest regressor, Gradient Boosting regressor, Light gradient regressor , Extra Gradient Boosting regressor and ADABoost Regressor. Then we use hyper parameter tuning on the best performing model to improve the accuracy further using the cross validation method on the parameter pairs.

## 4.Data Research:

### 4.1)Describing and Processing the data set:

In this data set we have features relating to listing descriptions, host description, neighbourhood of the listing, some important features regarding reviews. There are nearly 72 columns and close to 39,000 rows in this data set. There was a lot of clean up and processing that needed to be done as a lot of features were in the object data type even though they were integer and float values. There were also columns that had unnecessary characters which had to be removed. Some categorical features such as ‘amenities’ were converted into numerical features. And had to replace columns that had true(t) or false(f) values with 0’s and 1’s. There were a lot of missing values in the data set which were filled with the median values(for example for bedrooms,bathrooms etc) and with a false value for feature which describes if host is super host or not.

## 5.Analysis Strategy:

For visualization we use various scatter plots, bar plots, histograms and correlation heatmaps to find the correlation between different features.We will use the metrics MSE,MAE,RMSE and most importantly R-squared value to compare the efficiency of the model. After finding the best performing model we shall perform hyper parameter tuning to improve the accuracy more.To determine the accuracy of the model, model.accuracy() is not pragmatic to a consumer.As the prices fluctuate with time, (for example, price hikes on major holidays), It is more practical to bucket the prices and score the model using those buckets.We have used 35$ as the bucket size.

## 6.Analysis Code:

### 6.1)Exploratory Data Analysis:

After cleaning up the data and filling the null values we shall perform some exploratory data analysis to understand more about the data set. First we plot a scatterplot of different listings in different neighbourhoods according to latitude and longitude(refer figure.1). We cannot really tell where the majority of listings are as the plottings are overlapping each other. So we plot a simple bar graph(figure.2) describing the number of listings according to an area. In this we see that most of the listings are present in the Manhattan region followed by the Brooklyn region. Next we again plot a scatter plot of the listings based upon the latitude and longitude but this time the colour changes according to the price(figure.3-1), we see that this scatter plot is not accurate, maybe due to outliers. So we plot a simple histogram of the price feature(figure.4-1) and we find out that there are outliers in the data. We remove the outliers using the inter quartile range method and we plot the price histogram(figure.4-2) and the scatter plot(figure.3-2)again. We see that the scatter plot is more accurate and readable now.We plot a heat map to see the correlation between different features(figure.5). We can see that some features like reviews_ltm, reviews_l30d and review_scores have too much correlation. We will create a single feature for reviews and we will drop all the other columns. We do the same with other features as well. We can also see that there is more correlation between price and the accommodates, bathrooms_text, bedrooms and beds columns. We also plot the number of accommodations based upon the different room types of the listings(figure.6).

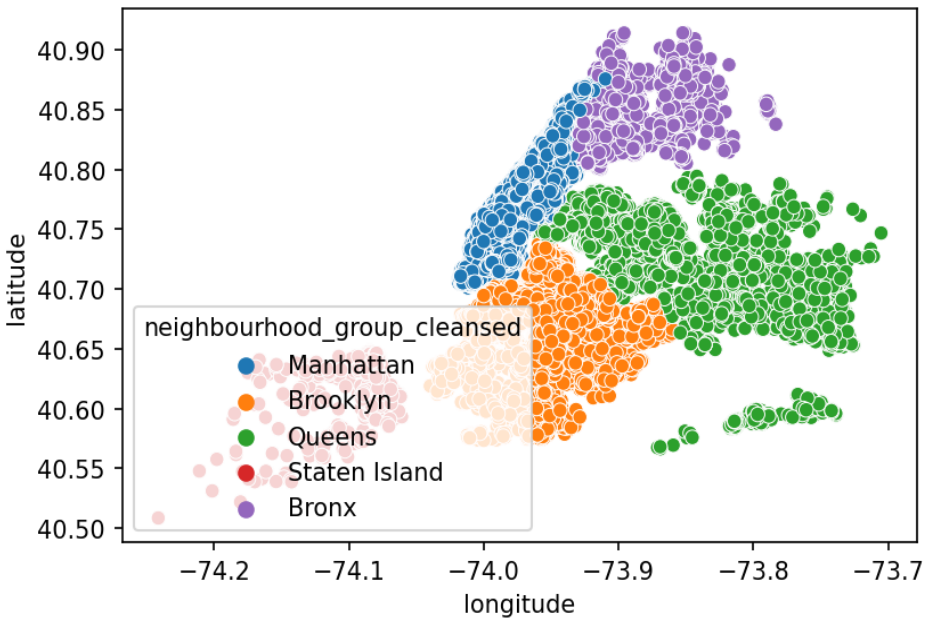


Figure.1

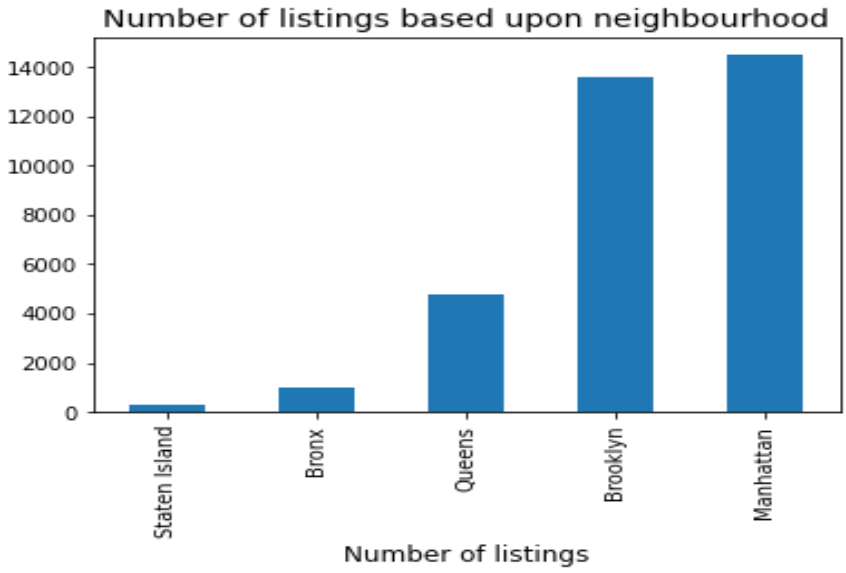


Figure.2

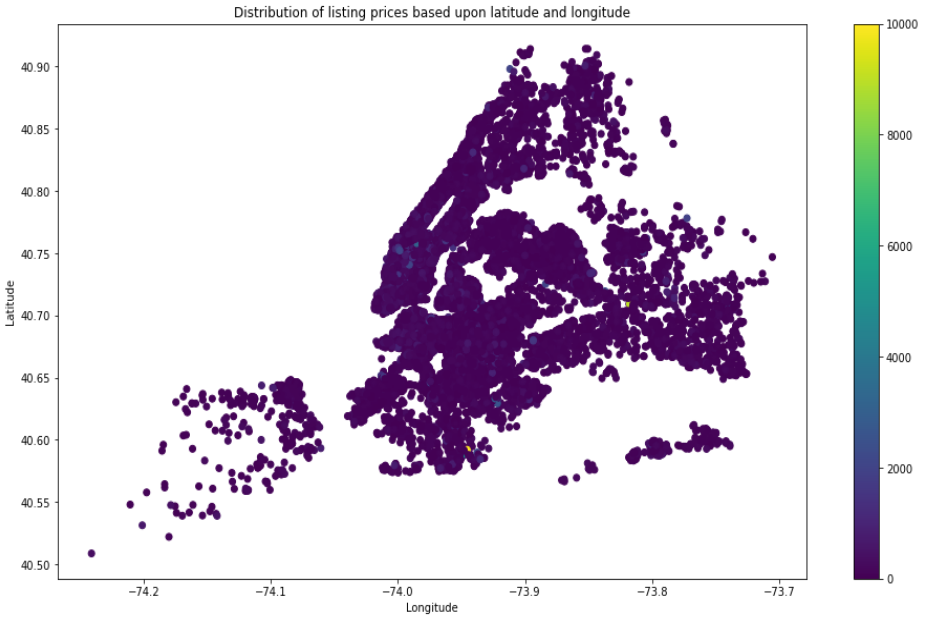


Figure.3-1

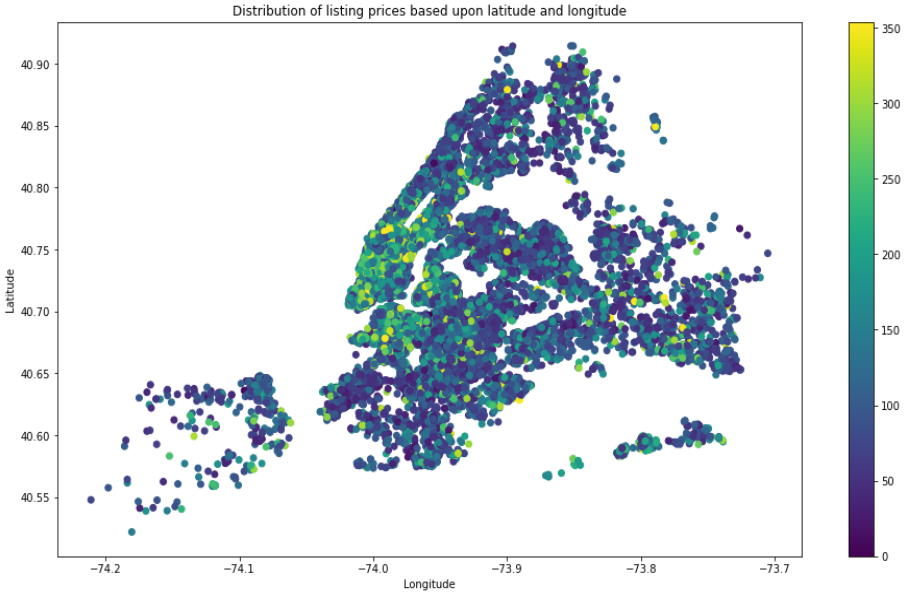


Figure.3-2

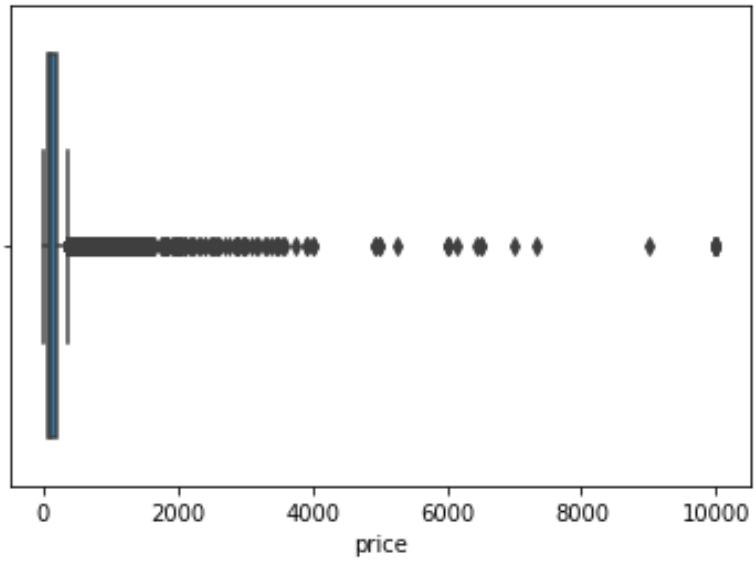


Figure.4-1

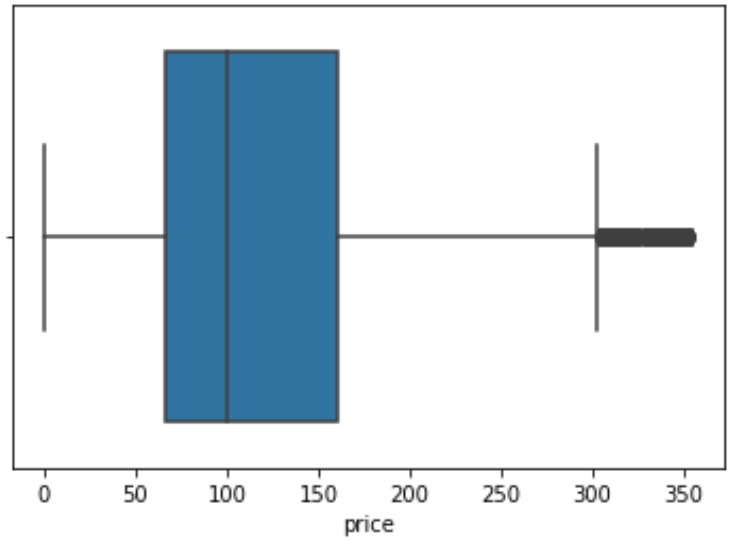


Figure.4-2

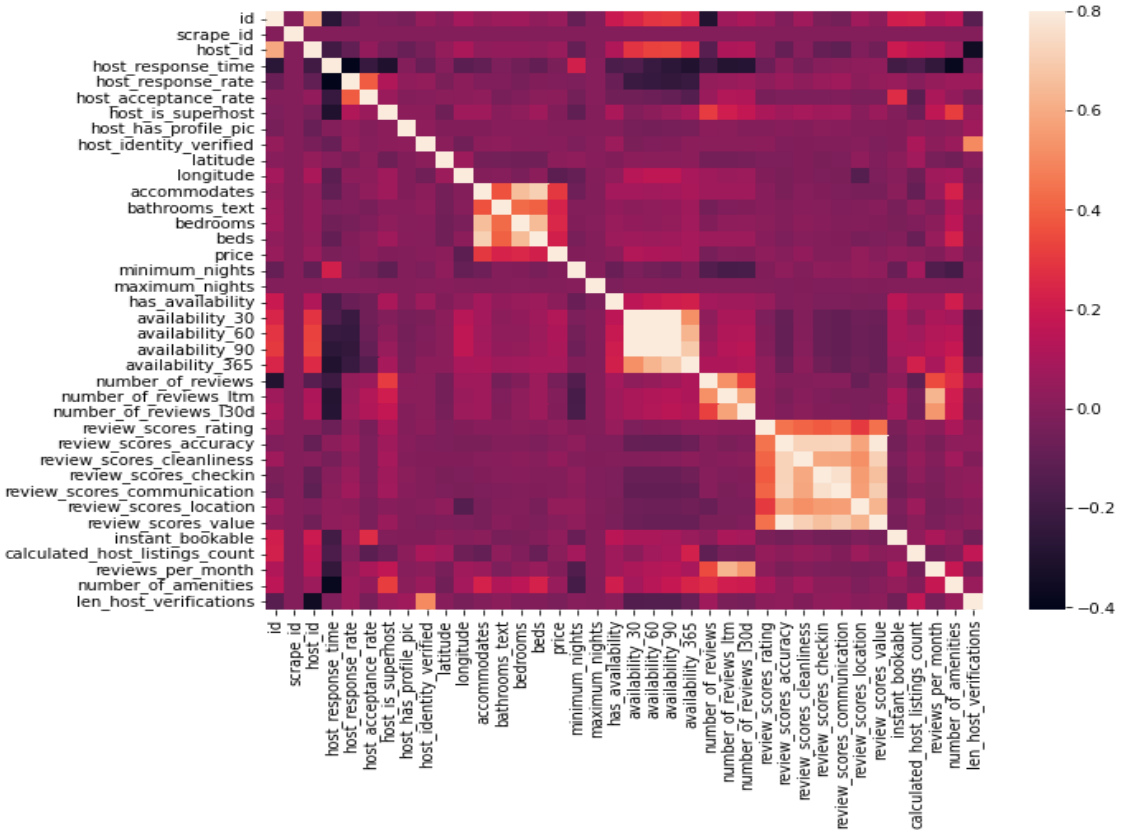


Figure.5

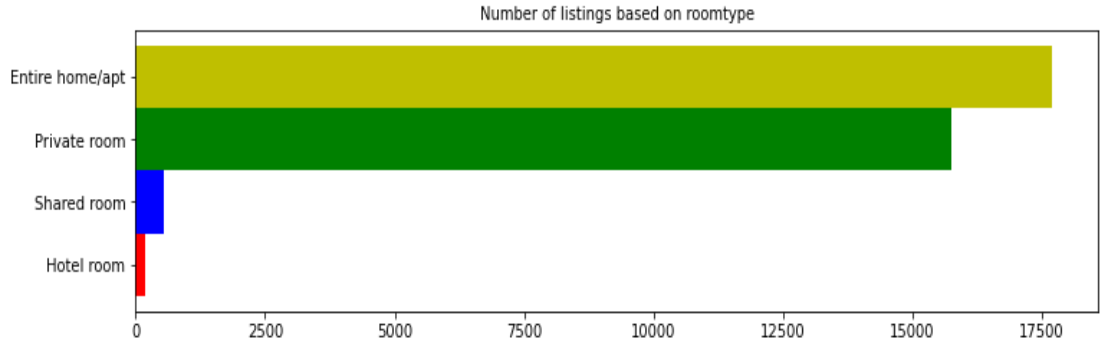


Figure.6

**6.2)Encoding Categorical Features:**

Price of the AirBNB listing will depend on the neighborhood it is in and the type of the room.But these columns are categorical values not numerical.So we use the one hot encoding method to create various new features.After creating the new features we drop the "neighbourhood_cleansed" and "room_type" columns.

**6.3)Performance of the models:**

| **Model** | **MAE** | **MSE** | **RMSE** | **R^2** | **+/-35 accuracy** |
|---|---|---|---|---|---|
| Linear Regression | 36 | 2434.94 | 49.35 | 54% | 62.61% |
| Random Forest | 30 | 1826.66 | 42.74 | 65.5% | 69.79% |
| Gradient Boosting | 31 | 1891.18 | 43.49 | 64.28% | 68.63% |
| Light GB | 32 | 1945.21 | 44.14 | 63.26% | 68.2% |
| Extreme GB | 29 | 1728.60 | 41.58 | 67.35% | 70.95% |
| AdaBoost | 44 | 3043.15 | 55.16 | 42.53% | 46.09% |
| KNN | 49 | 4439.81 | 66.63 | 16.15% | 48.06% |

Out of all the models we notice that the Extreme Gradient Boosting Model is the best performing model.

**6.4)Hyper Parameter Tuning:**

To increase the accuracy of the model we perform hyper parameter tuning on the model, here in our case XGB Regressor model. We use the Bayesian Optimization method to tune the hyper parameters and we run the fitting and predicting on the data set again. After the hyper parameter tuning the updated metrics are:

MAE: 29

MSE: 1760.62

RMSE: 41.96

R^2: 66.75%

Accuracy(+/-35): 71.16%

We can see that the accuracy has slightly increased.The increase in accuracy is less but the tuning is taking up a lot of space and computational time. Some of the cells for tuning are taking up to and more than 15 minutes to process.

## 7.Conclusion and Future Work:

Like we saw earlier, the Extreme gradient boosting regressor was the better performing model.We can notice that the prediction isn't accurate, with a score of **71% (+/-35)**.There will be a few reasons as to why the prediction might not be accurate and this might be due to several other features that are not part of the current data set we are using.

We could improve the accuracy of the models by having some extra features like sentiment of the reviews(as of now we only have the number of reviews and review scores etc). We can use NLP on the reviews to improve the prediction. Even though the listing is not close to the areas

with high pricing it might be close to various tourist spots or major destinations which will make it price higher, so we need a proximity feature which tells us the distance between the listing and the closest tourist destination or an important location. One more problem we might encounter is that even though two properties are in the same neighborhood the price of the listing might depend on how new or old the property is.A timeline of price trend could help predict the prices at major dates, such as holidays, new years, etc.

## 8.Work Planning and Organization of each team member:

Praharshita Kaithepalli: Data processing,EDA ,Model Development of linear regression and random forest regression, hyper parameter tuning.Fitting and prediction with the new parameters.
Nisarga Khairnar: Data processing,EDA ,Model Development of KNN,LGB,XGB,AdaBoost. Performance score coding.

## 9.Individual Contribution:

Dataset had columns which cannot be processed directly, thus to overcome this the columns had to be changed individually. For example, bedroom raw information contained string objects such as “ 1 shared” ,”2 private”,”1.5 shared”. These, string objects, had to be converted to the number of bedrooms and the type of bedroom. My contribution included creating new features such as host_importance and review_score. Host importance is a derived feature which is calculated by adding percentage of host response rate with the listing count and number of verifications of host, multiplied by the reviews given by customers to the host with respect to the communication. Host importance also includes whether a host is a superhost or not and has it’s identity verified or not. For better and practical accuracy, rather than using model.accuracy(), I

decided to bucket the prices. It would be more practical to give a range of how the price would be predicted as they change with respect to time. During Model development, due to the huge statistical noise in the dataset, I decided to test KNN regressor, with expectations of giving bad results as listings which have similar features had quite some difference in price. To verify that, KNN regressor did give the results as expected with an r2 score of 16 and +/-35 accuracy of 48%. After confirming the data noise, I decided to test gradient boosting algorithms as they would catch the variance of the feature set with respect to the output label better than previous models. The XGB model gave the best accuracy as seen, thus deciding to proceed with hyper parameter tuning. Using Bayesian Optimization algorithm, we got a robust model which gave little more error but a better accuracy.

## 10.Improving teamwork and collaboration:

Could have improved communication while working on the models. Splitting of work should have been done earlier as both the team members just kept going back and forth on the project.