Survey Analysis on Hotel Room Price Prediction in NYC

Hemant Kumar Das and **Mayank Lara**, Department of CSE, SUNY Buffalo.

***Abstract*—**In the age of data-driven solutions, the customer demographic and satisfaction attributes, such as neighbor -hood group, room type, longitude, number of reviews, play a major role that may enable the customer to decide the right hotel room for the value. Since there is no convenient platform as of now for the customers to check hotel prices in the entire city based on the specific requirements such as calculated host listings count and its availability around the year, this model would provide a platform to the customers to book hotel rooms based on their desideratum. Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019. This paper reviews diverse literature to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions. Several Machine Learning models have been implemented to analyze and predict the data such as Ordinary Least Square, Support Vector Regression, Random Forest Regressor, XGBoost Regressor, etc.

***Keywords*—**Hotel, Room, Price, Models, NYC

I. INTRODUCTION

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services, and much more. This dataset has around 45,000 observations in it with 16 columns and it is a mix of categorical and numeric values. For this purpose, the machine learning approach is applied which employs several regression algorithms for predicting the hotel room price based on the neighborhood group, room type, longitude, number of reviews. To address this, supervised learning algorithms as regression techniques are considered. A regressor maps input variables to the target class by considering training data. Regressors addressed in the project are described briefly. These regression-based predictions give us Root Mean Squared Error which is around 50$.

II.REGRESSORS

1) Ordinary Least Squares: **Ordinary least squares** (**OLS**) is a type of linear least-squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function of the independent variable.

2) RMSE: **Root Mean Square Error** (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how to spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

3) Decision Tree Regressor: Decision tree builds regression or classification models in the form of a tree structure. It breaks down a

dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**

4) Random Forest Regressor: Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

5) Support Vector Machine Regressor: Support Vector Machine can be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because the output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem.

6) XGBoost Regressor: XGBoost is a powerful approach for building supervised regression models. The results of the regression problems are continuous or real values. Some commonly used regression algorithms are Linear Regression and Decision Trees. There are several metrics involved in regression like root-mean-squared error (RMSE) and mean-squared-error (MAE). These are some key members of XGBoost models, each plays their important roles.

III.PROPOSED METHODOLOGY

The target of this study is to predict the hotel room prices in New York City based on various factors such as neighborhood group, room type, longitude, latitude, number of reviews. Identifying and eliminating unnecessary data columns and implementing machine learning models in this survey dataset will help us to investigate the impact of the neighborhood group, room type, longitude, latitude, number of reviews on the price. In this context, a dataset from Airbnb is processed that provides information regarding the describes the listing activity and metrics in NYC, NY for 2019.

A) PREPROCESSING STEPS:

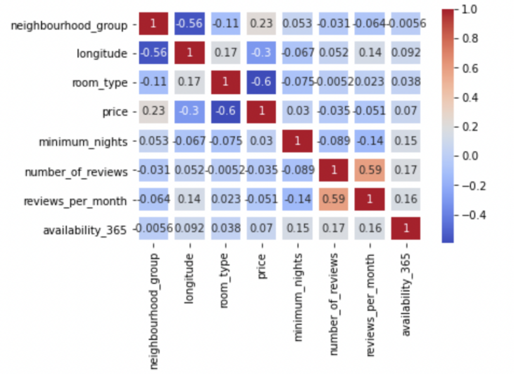
1. Checking and dropping duplicate rows helped in removing 1500 redundant rows.
2. Removing the old records based on the last review date.
3. Sorting the dataset based on the last review date, Considering only the records which are starting from the 2000 index, since the records before are quite old and futile.
4. Dropping all the unnecessary columns required to predict which helped in increasing accuracy.
5.Resetting the index, since the index before was jumbled and not in order, checking what all unique records are there for the neighbourhood_group column.
6. Categorizing the room type group from 0-2.
7. Filling all the reviews per month which is empty with 0, since the number of reviews that are empty refers to zero.
8. Calculation of Variance Inflation Factor to check for multi-collinearity.
9. Dropping highly correlated column 'latitude' which has high VIF.
10. Detecting and removing outliners using Z Score.

B) EXPLORATIVE DATA ANALYSIS:

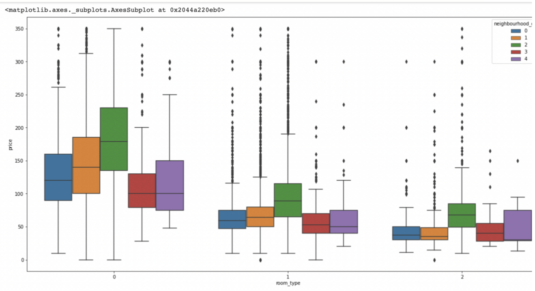
1. This heatmap gives us correlation values between every column of my dataset. If the value is very close to -1 or 1 we need to remove one of the columns of the two correlated columns(other than target variable 'price'). If we have highly correlated data then my linear regression (which I am planning to do) will have incorrect values in coefficients. If the correlation value is near 0 then we have un-correlated data which is desired for features other than the target variable. Here, we do not have highly correlated

data other than the correlation with 'price'. So all columns will be considered for model building.

```
<matplotlib.axes._subplots.AxesSubplot at 0x204490b79a0>
```

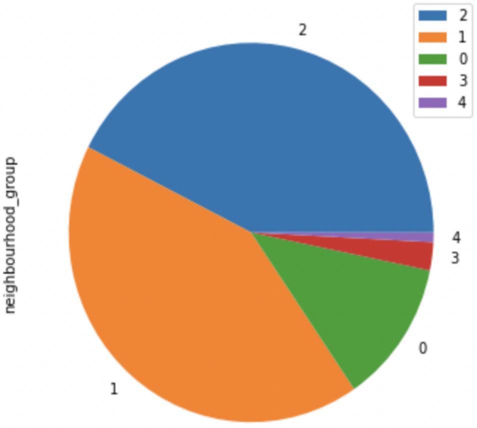


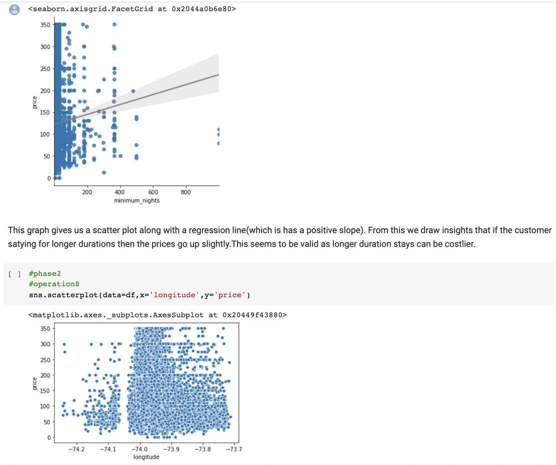
2.Box Plot: This is a box plot that gives us information about the median price and inter-quartile range of the different types of rooms as well as the neighborhood groups they are in. We can see that the median price for Neighbourhood 2 ('Manhattan) is greater as compared to other neighborhoods of NYC. We can also see that the median price for room type 0 (Entire home/apt) is greater than other room types. So Entire 'home/ apt' in 'Manhattan' will most likely be a costly deal.



```
<matplotlib.axes._subplots.AxesSubplot at 0x2044a220eb0>
```

3.Pie Chart: Queens": 0, "Brooklyn": 1, "Manhattan":2, "Bronx": 3, "Staten Island":4. From this we can see that there are very few rows with 'Bronx' and 'Staten Island' in 'Neighbourhood_group' feature. We can conclude that the number of hotels in 'Bronx' and 'Staten Island' is fewer as compared to 'Manhattan','Brooklyn' and 'Queens'. Hotels are mainly located in 'Manhattan' or 'Brooklyn'.





```
<seaborn.axisgrid.FacetGrid at 0x2044a0b6e80>
```

This graph gives us a scatter plot along with a regression line(which is has a positive slope). From this we draw insights that if the customer satying for longer durations then the prices go up slightly.This seems to be valid as longer duration stays can be costlier.

```
[ ] #phase2
    #operation8
    sns.scatterplot(data=df,x='longitude',y='price')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x20449f43880>
```

The above represents a scatter plot between longitude and price.

C) IMPLEMENTED REGRESSORS

As the data is ready to be trained, a total number of 5 regressors are used. *Ordinary Least Squares, Decision Tree Regressor,* ***Random Forest Regressor,*** Support Vector Machine Regressor, XGBoost Regressor are trained, and their RMSE has been noted.

For maximizing the performance of these models, default parameters may not be sufficient. Adjustment of these parameters enhances the reliability of this model which may be regarded as the optimized one for identifying as well as isolating the redundant response.

D) *EVALUATION METRICS*

To evaluate the performance of the model we need to have some metric that will measure the output of the model with the ground truth values. For this purpose, the following metrics are taken into consideration to identify the best relevant problem-solving approach. Based on the data provided to the models and trained models, the following details in the table will illustrate the performance of the model over the test data.

Root Mean Square Error *(RMSE) is the standard deviation of the residuals (prediction errors).*

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

OLS	Decision Tree	Random Forest	SVR	XGBoost
59.89	70.08	50.24	52.86	49.33

V. CONCLUSION

This paper has revealed the overall price prediction of the hotel room price by taking into account the factors such as neighborhood group, room type, number of reviews, availability 365, latitude, longitude, etc. We find that if the customer is staying for longer durations then the prices go up slightly. This seems to be valid as longer duration stays can be costlier. Also, the costliest rooms are available when the longitude is between -74 and -73.9. This area has the costliest places to stay. If you check the longitude of Manhattan it is -73.97 and Brooklyn has -73.94. We conclude that the costliest hotels lie in Manhattan and Brooklyn. For handling the analysis among them, several machine learning algorithms are proposed as countermeasures in our approach. The supervised learning method is used to identify the use of several regressors for finding the hotel room price. Experimental results indicate that XGBoost outperforms overall other regressors by achieving an RMSE of 49.33 which is much better than the existing methods.

VI. REFERENCES

1. https://www.geeksforgeeks.org/root-mean-square-error-in-r-programming/
2. https://machinelearningmastery.com/xgboost-for-regression/
3. https://en.wikipedia.org/wiki/Random_forest.