New York City Airbnb Rent Prediction

Pooja Yadav
Dept. Computer Science
San Diego State University
San Diego,California,United States
pyadav5723@sdsu.edu

Shweta Shete
Dept. Computer Science
San Diego State University
San Diego,California,United States
sshete8436@sdsu.edu

***Abstract*—Airbnb is an online housing platform that provides users to list, discover and book accommodations for their users all across the world. Hosts can offer their property spaces to guests for short or long periods.Airbnb's listings offer a huge range of options ranging from community bedrooms to luxury housing, all in one platform. Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. When traveling to other cities or countries, accommodation is the principal necessity. One of the main things in choosing accommodation is the reservation price and the locality. What are the main factors which affect reservation prices? Which host is the busiest or most reviewed? What other factors affect the price. Locality, amenities, reviews, or any other, we try to analyze that using Airbnb dataset. This dataset describes the listing activity and metrics in NYC, NY, for the year 2019. The objective of this paper is to predict the rental price for an Airbnb accommodation. To achieve this a predictive model is built using the regression techniques, and to pick the best performing model a comparative analysis of their performance is measured using different performance metrics. The model accuracy is measured using performance metrics such as Root Mean Squared Error (RMSE), R-squared score (R2) metric, and Variance Score.**

***Index Terms*—Airbnb,prediction,regression,metrics,hosts.**

I. Introduction

Airbnb started in 2008 when two designers who had space to share hosted three travelers looking for a place to stay. Now, millions of hosts and travelers choose to create a free Airbnb account to list their space and book unique accommodations anywhere in the world. Airbnb experienced hosts share their passion and interest with the travelers and the locals[2]. Now, Airbnb has been the first choice for individuals and groups who are looking for lodging service and tourism experience other than hotels. Tourists are mainly motivated to book Airbnb accommodation because of its low cost, convenient location, and household amenities[1]. Due to which there arises a need to determine the rent for an Airbnb accommodation. Deciding the price of an Airbnb accommodation is important for the apartment owners as it is a deciding factor on the number of customers. On the other hand, customers need to evaluate based on the decided price. This paper aims to develop a reliable rent prediction model using the regression models such as linear regression and XGBoost regressor. The data used for prediction is of New York City of the year 2019. Features such as types of rooms, number of reviews, location, and availability are analyzed to form the prediction model. It also shows the understanding of the expensive neighborhoods of NYC city, availability of hosts, using visualization techniques. The prediction is based on the top five neighborhoods of NYC City. The Linear Regression model and XGBoost give a nearly accurate prediction of the above data.

II. Approach

A. Dataset

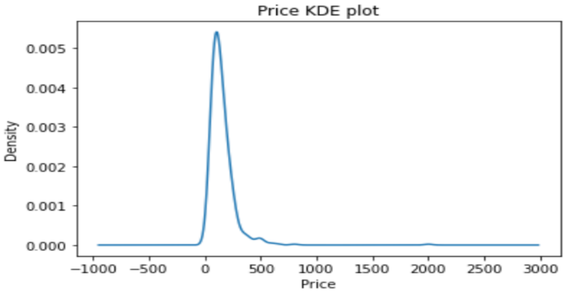
The Dataset used was from kaggle.It has sixteen columns and 48895 entries

B. Data Cleaning

we first identified the unique neighborhood group and room-types from the dataset,there are five neighborhoods and three types of rooms available.Then handled all the missing values by replacing them with zero's as they do not serve any important role in our prediction

C. Data Visualization

To check the most common prices for Airbnb we check by plotting prices against its density by using the KDE plot. A kernel density estimate (KDE) plot is a method for visualizing the distribution of observations in a dataset,analogous to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions[3]. [htbp]



fig(a): KDE plot of price

Then we check the availability of hosts and its maximum listing by using barplot. A bar plot represents an estimate of central tendency for a numeric variable with the height of each rectangle and provides some indication of the uncertainty around that estimate using error bars. Bar plots include 0 in the quantitative axis range, and they are a good choice when 0 is a meaningful value for the quantitative variable, and you want to make comparisons against it[4].

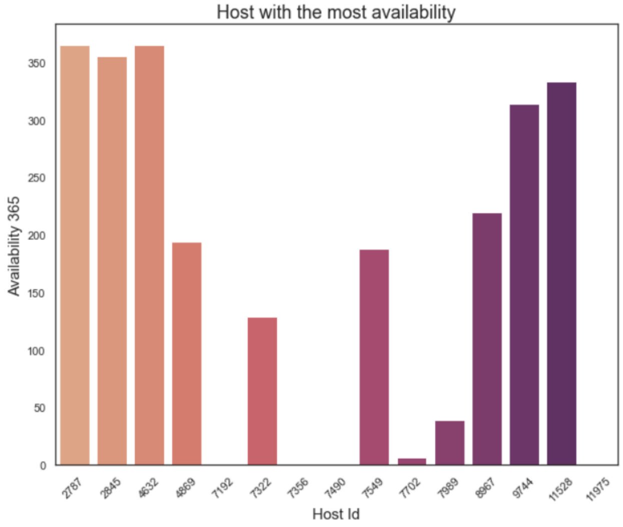


Fig. 1. fig(b):Maximum Hosts Availability

To determine the price prediction we need to evaluate the rooms based on neighborhoods so we find the most expensive neighborhoods using the violin plot. A violin plot plays a similar role as a box and whisker plot. It shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared. Unlike a box plot, in which all of the plot components correspond to actual data points, the violin plot features a kernel density estimation of the underlying distribution[5].

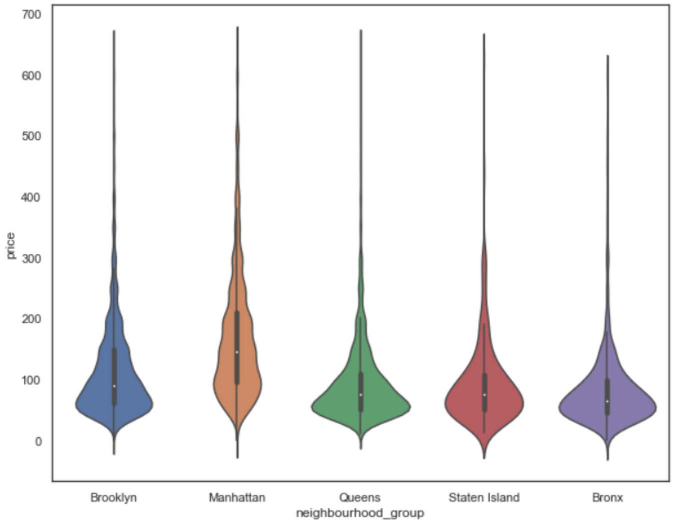


Fig. 2. fig(c):Neighbourhood Group Prices

We have applied one-hot encoding on the neighbourhood and room-type to convert them into binary values. In one hot encoding, the features are encoded using a one-hot (aka 'one-of-K' or 'dummy') encoding scheme. This creates a binary column for each category and returns a sparse matrix or dense array (depending on the sparse parameter).[6].

We also drew a correlation heatmap for all columns using pearson's correlation coefficient.Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.[7][8]

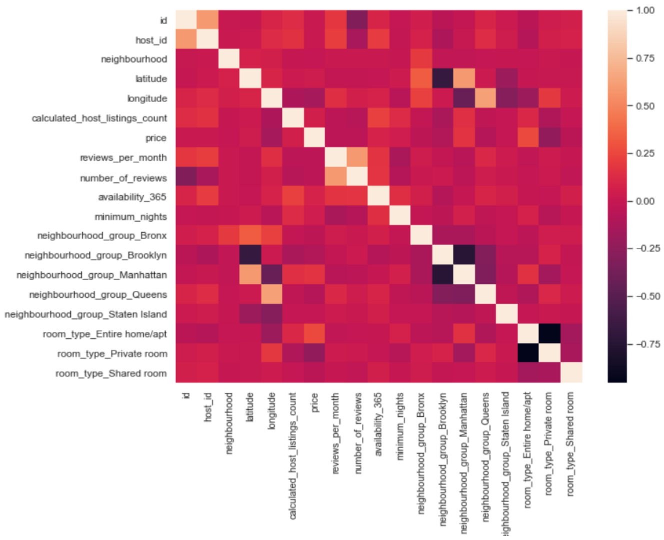


Fig. 3. fig(d):Heatmap

III. Modelling

A. Linear Regression

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).[10] To achieve the rent prediction goal we applied Linear regression model on the trained data after performing all the above steps.

B. XGBoost Regression

To compare between models and choose the best regression model we also used XGBoost regression model on our trained data. XGBoost is an efficient implementation of gradient boosting that can be used for regression predictive modeling.[9]

IV. Evaluation

For Evaluation of the model, we use Root Mean square Error (RSME),Mean Squared Error(MSE),R2 Score and Variance Score obtained from both the regression techniques. For

XGBoost:RMSE: 61.178859 R2 Score:0.9073260149524471 Variance score: 0.91 From the above observation, we can see that XGBoost gives a better accuracy of rent prediction

V. Conclusion

From the above observations given in the paper we can conclude that the XGBoost Regressor gives an Accurate Rent prediction Model with an accuracy of 90

References

A. Sites

[1]https://hospitalityinsights.ehl.edu/
[2]https://www.airbnb.com [3]https://seaborn.pydata.org/generated/seaborn.kdeplot.html
[4]https://seaborn.pydata.org/generated/seaborn.barplot.html
[5]https://seaborn.pydata.org/generated/seaborn.violinplot.html
[6]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
[7]https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/
[8]https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/
[9]https://machinelearningmastery.com/xgboost-for-regression/ [10]https://en.wikipedia.org/wiki/Linear-regression

References

[1] K. Shah, H. Shah, A. Zantye and M. Rao, ”Prediction of Rental Prices for Apartments in Brazil Using Regression Techniques,” 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 01-07, doi: 10.1109/ICCCNT51525.2021.9579796

[2] J. Dhillon et al., ”Analysis of Airbnb Prices using Machine Learning Techniques,” 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 0297-0303, doi: 10.1109/CCWC51732.2021.9376144.

[3] S. Yang, ”Learning-based Airbnb Price Prediction Model,” 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), 2021, pp. 283-288, doi: 10.1109/ECIT52743.2021.00068.

[4] Rezazadeh Kalehbasti P., Nikolenko L., Rezaei H. (2021) Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. In: Holzinger A., Kieseberg P., Tjoa A.M., Weippl E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2021. Lecture Notes in Computer Science, vol 12844. Springer, Cham.