Predicting prices of New York City's AirBnb

Tanvi Nautiyal
MSc Data Analytics
National College of Ireland
x19145381

I. Abstract

AirBnb has gained its popularity within past few years and customers now prefer AirBnb more than any hotel while booking their stay. Airbnb not only provide cheap stays but also gives quality and 5-star experience to its customer. They're very responsive with their feedback system and implement the things which customers find missing in order to maintain its quality. New York one of the most visited city of the world has seen a huge spike when it comes to tourism where AirBnb has played a vital role handling the hospitality sector. There are tons of AirBnb present in the New York area with variety of rates and services. The main goal of this research is to create a price prediction model that will help the AirBnb and the customers as well to have a bird's eye view about the Airbnb and its prices all throughout the year. For this research certain machine learning models are considered out of which Multi Linear Regression is the best suited one for the current dataset, as it will give the best price predictive model for the selected AirBnb dataset

Keywords— Airbnb, Price Prediction, Machine Learning, Multi Linear Regression, CRISP-DM

II. Introduction

Over the last few years, people have moved around more frequently as they used to do before, making the market for the hospitality sector is grow. Short term facilities come to play here to satisfy this ever-increasing demand. Most hotels have begun to increase their rent after seeing a sharp rise in the tourism domain. Airbnb, an online marketplace where guests who are looking for accommodation can share experiences and homes with guests who have extra rooms to rent [7].

Frank Yang et al. in [7] states that it is the host who decides the price of the property and how it dynamically changes. However, by comparing with other properties in the same here or during the holiday season, Airbnb does provide suggestions for these hosts on how they can increase or decrease the rate of the property.

Airbnb's is designed in such a way that private owners of any listed properties can supply their property within 191 countries all around the world. The site operates over a total of 4 million Airbnb listings across the 191 countries, and more than 2 million users use the Airbnb to book their stay or can rent their property as a host.

The Airbnb platform also enables organizing of certain activities for individuals including "host guests."The majority of revenue generated by Airbnb is from members of the Host.Host members are responsible for posting and providing information for the properties / events.The traveler will be able to view the huge house listings using the Airbnb website.Most members can book a trip to nearly anywhere in the world

When it comes to hospitality and tourism, Airbnb is considered one of the most promising markets, because it offers not only inexpensive but also comfortable accommodation. Since its launch in 2008, Airbnb has risen steadily in terms of revenue growth and its broad variety of services and people like Airbnb's more than any other short-term rent. There are now over 100 million Airbnb users in many countries making the conventional hospitality industry a major disrupter. Airbnb receives revenue from both hosts and visitors, where 3% of the reservation value goes to the hosts arranging the stay and customers are charged between 61.2% of the reservation. Airbnb produces tons of data as rental ecosystem. New York, along with several other cities, is one of the popular markets for Airbnb. It can be very difficult to rent an Airbnb in New York City because one must work out the best rates, location and several other considerations while booking. Since New York is one of the most expensive cities, one must be careful which Airbnb to choose.

III. Literature Review

Zhihua Zhang et al [1], implemented general linear model and geographically weighted regression to check the significant factors or variables impacting the price of Airbnb in Metro Nashville Tennessee area. From the outcome in [1], geographically weighted regression outperforms the general linear model that depicts that the price of Airbnb listing varies with the location. The Airbnb's located in the city centre tends to have high prices than the ones situated outside the city. The result obtain provides a better view over the market situation and can help stakeholders in developing a better business strategy.

Javier Gutierrez et al [2], analysed the spatial distribution of Airbnb listings, hotels and sightseeing spots in Barcelona region. For this study geolocated data sources are used and through this study Airbnb focused in the city centre region majorly in the residential areas of the city. Multi Linear Regression model is implemented that depicts that the Airbnb listings having close vicinity to the sightseeing places profits more than the hotels in the same region.

In [3] Dan Wang and Juan L Nicolau, studies the Airbnb listings prices of 33 cities. In this paper price is considered as the response variable and how is being affected with five

sets of explanatory variables and which independent variable affecting price variable the most. For analysing this Linear QR models and Linear OLS regression models are implemented. The result obtain from both the models defines that the explanatory variable "superhost" leads to the higher prices with 8.73% of the actual price, whereas profile picture of host shows a reduction of 10.89%. Verified Hosts yields a positive response and price increases by 8.94%. Countries like USA, Australia, Germany here price is significantly high than the countries like Italy, Canada,Ireland.

Hoormazd Rezei et al [4], implemented machine learning models and sentiment analysis for predicting the price of Airbnb listings in New York City Region. To achieve this methods like Linear Regression,Support Vector Regression, K-means Clustering and Neural Networks are executed in this study for creating a price forecasting model. Here SVR provides the best results as they're having higher R square value of 69% and MSE value of 0.147.

Emily Tang and Kunal Sangani [5], studies the data of Airbnb Listings in San Francisco region and predicted which neighbourhood is best and created a price forecasting model for achieving this Support Vector Machine classifier using feature extraction. Here Up state neighbourhoods are having higher prices.

In [6] Libuse Haubeltova, predicted the prices of Airbnb Listings in Berlin region. For that Linear OLS Regression model is implemented keeping price as a response variable and rest other variable as explanatory ones. Here the null hypothesis is that sharing economy variable holds no statistic significant on price and hence it is rejected. If a host is having a status of super host then there can be an increase in the average price by 12%.

In[8], describes how implementation of machine learning algorithms such as Multi Linear Regression, Lasso Regression,Ridge Regression, Elastic Net Regression, Ada Boosting and gradient boosting method can impact the variance in housing price prediction. A comparative analysis is made, out of which gradient boosting method gives the best accuracy among all. Here accuracy is compared with the values obtained from RMSE (Root mean square error) and MSE (Mean square error).

Kevin Han et al [7], studies various machine learning algorithms such as Linear regression, ridge regression, random forest regression, support vector regressor, neural network and gradient boosting on the Melbourne Airbnb. After implementing these many models, a comparative analysis has been made and Gradient boosting method when all features are included, gives the best performance followed by random forest regression as the second best.

Tao Yang et al. [9], created a price prediction paper for holiday rental websites where rental property prices typically used for travel or holiday purposes are predicted using different regression models. In addition, they used a multi-scale method of propagating affinity, which is basically the clusters where the unwanted data is extracted from the dataset.

Huija Yu, Jiafa Wu [10],works on predicting the housing prices of the real estate region by implementing regression and classification techniques. The primary purpose of this research was to carried out a comparative analysis and consider the one which gives better result. PCA technique is implemented to enhance the accuracy of the research. A comparision is made with the different algorithms of SVM where SVM using gausian kernel outperform the SVM using linear kernel and achived rmse value of 0.5271 and final accuracy of 69.1%.

Yixuan Ma et al. [11], presented their work by predicting the rental prices of warehouses present in China by using machine learning approach. After running machine learning algorithm they concluded that the prices of warehouses are directly proportional to the land and location. If the location where the warehouse is present is popular then the rental prices of those warehouses will be higher. To predict this they have used the random forest regressor acheiving a correlation matrix for test set of 0.57.

IV. Methodology

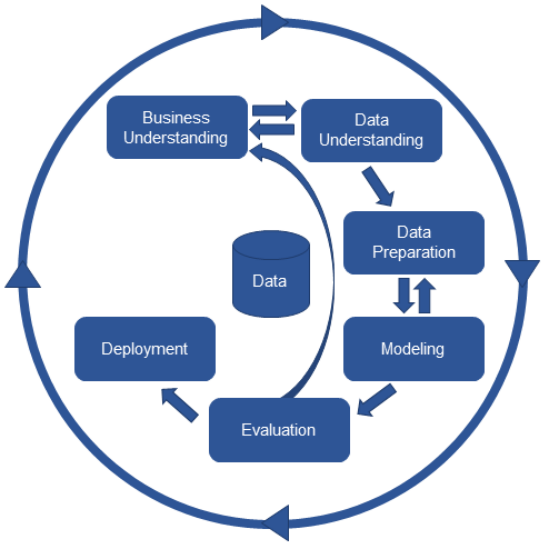

Fig. 1. Methodology

The methodology selected for this research will provide a brief idea of the business understanding of this research. Not only that methodology will give a brief understanding how machine learning model is implemented and how exactly a research is being carried out and if there are any errors showing up one can always trace back the root cause with the help of this methodology.

A. Business Understanding

In predictive analytics it is very important to have a proper business understanding about the data. For NYC AirBnb, it manages the two side of it. First one is the customers and the other one is the hosts. A valuable feedback form both the side is very important for an AirBnb to function. To stay this popular in the market AirBnb has provided a huge rating and feedback system where the customer can rate their stay

and it can help AirBnb to improve if there arises any room for improvement also from the host perspective it can keep a check how well that host's performance is going and if there is any sign of improvement it will directly inform the host so that there won't be any negative feedback. Also, if the performance of host is increasing it will allocate host as superhost which is most preferred by the customers and that's how AirBnb monitors its business model.

B. Data Understanding

The following data considered for this research is NYC AirBnb Open Data which consists of 48,895 rows and 16 columns in total keeping a track of all the AirBnbs in New York region in 2019 and is available from Kaggle repository,[8].Following is the data description of the NYC AirBnb open dataset:

Variable	Description
ID	Row ID
Name	Name of the AirBnb
Host_id	ID of the host
Host_name	Name of the Host
neighbourhood_group	Zone where Airbnb is falling
neighbourhood	Area where AirBnb is present
latitude	Latitude of the AirBnb
longitude	Longitude of the Airbnb
room_type	Type of room Airbnb is providing
price	Price of the Airbnb property
minimum_nights	Minimum stays in Airbnb property
number_of_reviews	Number of reviews given to the property
last_review	Date when the last review given
reviews_per_month	Average of review given on monthly basis
calculated_host_listing_count	Number of property each host have
availabilty_365	Availability of the AirBnb property

Fig. 2. Data Description

The figure below shows the overall structure of all the AirBnbs spread all across New York region.The following map gives the brief view of the neighbourhood in New York and one can get an idea where the maximum number of Airbnbs present and which is the most popular zone.

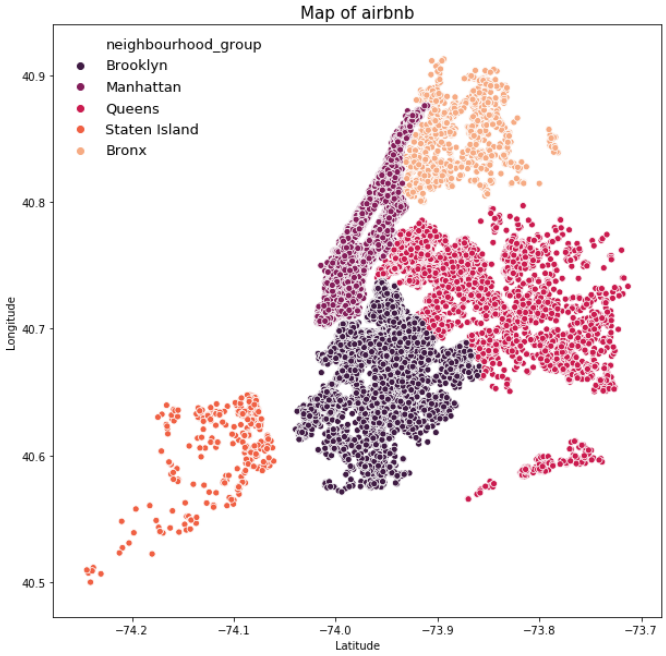

Fig. 3. Airbnbs all across New York

C. Data Preparation

Data Preparation is one of the most important step before running any machine learning model as it will make data more accurate and remove the unwanted values present in it . When data is passed for the regression or classification it should be structured in such a way it will get easier for the model to interpret. For this dataset certain pre processing steps are implemented to make the data more accurate. The very first step is checking the missing from the figure below it can be seen that there are missing values present for to overcome that those rows are dropped from the data as they do not contribute much.

```
Null values in Airbnb dataset:

id                                  0
name                               16
host_id                             0
host_name                          21
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                     10052
reviews_per_month               10052
calculated_host_listings_count      0
availability_365                    0
dtype: int64


Percentage of null values in review columns:
20.56 %
```

Fig. 4. Checking for null values

And shall convert room type and neighbourhood group into categorical variables. Once this is done the data is further split

into test or train data in a ratio of 80:20 and it is now ready for the next phase.

D. Modeling

Multi linear regression also known as multi regression defines a relationship between explanatory variables(independent) and response variables(dependent). It can be defined from the following equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

For this research Multi Linear Regression is one of the best suited machine learning algorithm when it comes to predicting prices. Here, price is considered as a dependent variable and room type,neighbourhood group,minimum night and availability 365 are considered as independent variables.

E. Evaluation

To check the performance of the machine learning model their accuracy is checked. But since we're using Multi Linear Regression of this research hence we can check the performance on the basis of R-Square value. In this model since the price variable was skewed therefore model is run with and without log function to check the value of R-square. When it is run without log function then the values comes out to be 0.4171 and with log function it comes out to be 0.5367. Later RMSE,MSE and MAE value are calculated for the model.

```
Call:
lm(formula = log(price) ~ neighbourhood_group + latitude + longitude +
    room_type + minimum_nights + availability_365, data = train_airbnb)

Residuals:
    Min      1Q  Median      3Q     Max
-2.93143 -0.28313 -0.03116  0.24682  2.41854

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                   -1.696e+02  7.128e+00  -23.788  < 2e-16 ***
neighbourhood_groupBrooklyn   -3.390e-02  1.954e-02   -1.735   0.0828 .
neighbourhood_groupManhattan   2.471e-01  1.774e-02   13.927  < 2e-16 ***
neighbourhood_groupQueens      7.874e-02  1.884e-02    4.179 2.94e-05 ***
neighbourhood_groupStaten Island -7.452e-01  3.632e-02  -20.515  < 2e-16 ***
latitude                      -5.724e-01  6.975e-02   -8.208 2.34e-16 ***
longitude                     -2.674e+00  8.014e-02  -33.374  < 2e-16 ***
room_typePrivate room         -7.186e-01  4.843e-03 -148.386  < 2e-16 ***
room_typeShared room          -1.126e+00  1.522e-02  -73.997  < 2e-16 ***
minimum_nights                -1.939e-03  1.149e-04  -16.878  < 2e-16 ***
availability_365               5.182e-04  1.827e-05   28.369  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4238 on 33348 degrees of freedom
Multiple R-squared:  0.5367,    Adjusted R-squared:  0.5366
F-statistic:  3864 on 10 and 33348 DF,  p-value: < 2.2e-16
```

Fig. 5. R-Square Value

F. Deployment

Once the model is executed and price prediction model is created then the same can be implemented on AirBnb's end as it can give a brief detail on the New York Airbnbs and how they are fluctuating all throughout the year. By implementing this model one can get an overview on the prices of Airbnb and when they are varying and can book their stay as per that.

V. Quantitative Result

The Multi Linear Regression implemented in this research achieved total accuracy/R-square of 53.67% which is quite good keeping in mind that the price of AirBnbs constantly changing. Here the neighbourhood group is sub categorized into 5 neighbours as they were given quite a good significance rather than calling them all together. The value of RMSE is increased when log function is implemented in Multi Linear Regression. For multiple groups of society such as customers, homeowners, real estate agents etc., this knowledge can be very helpful. This will help customers to determine which areas they will be looking at based on their needs and house owners should determine on the prices they put on the listings taking all these considerations into account.

VI. Qualitative Result

A predictive model was necessary it order to keep a check on the prices of Airbnbs, it will not only help the customers but the AirBnb itself to enhance their business as they will be knowing when to increase the price to attain a maximum profit.AirBnb's main partners are the hosts and the customers so in order to maintain a proper balance it has to check both it ends. From host's perspective it will check whether the host is providing proper and correct accommodation and to it customers and that the customers have made the right decision choosing the property.Also the hosts are alloted a title of superhost if certain number of customers liked the property and have given a good feedback. This title can help the hosts to get more and more customers and can profit AirBnb too. On the other end customers feedback are very important if any Airbnb need to make some improvements and Airbnb has got a really good response rate and they immediately implement the thing that customers mentioned it to be missing.

This extremely fast response rate of Airbnb makes it stand out from the rest of the hospitality sector and attracts more and more customers. In order to maintain the business Airbnb keeps a track on its feedback system and maintain both the ends also the mobile application provided by Airbnb is very user friendly and can be easily understood by the users and they can book their stay without any hassle. In additional to that Airbnb also provide the similar property options to its customers if they former place is occupied or is not available so that the customer can book the other almost same option with no increase in price. All these facilities are provided by Airbnb to keep their business intact and to maintain a spot in the market beating all it competitors.

The values obtained from the model shows that New York Airbnb still needs to improve their structure in order to attain better R-Square value and their is still a room for improvement.When it comes to business prespective, more business strategies are needed to minimize the current error and maximize the profit and also provide the efficient and

budget friendly services to its customers in order to stay this popular in the market.

VII. Conclusion

Airbnb is a private, multinational business providing hospitality services on-line marketplace website. Airbnb can be reached via their official website or through an app. Airbnb users may use the site to arrange or sell homestay lodgings or tourism experiences. For this company to function it is essential to maintain the price prediction model as its keep on changing all throughout the year. To create a price prediction model Multi Linear Regression is called and it attained R-Square of 0.5367 when log function is used else it was scoring R-Square of 0.4171, which shows that our New York Airbnb data need a room for improvement. In near future other machine learning algorithms can be implemented such as regression tree or decision tree and then their performance is checked and compared and the one giving the best performance can be consisdered for this dataset.

References

[1] Z. Zhang, R. Chen, L. Han and L. Yang, "Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach", Sustainability, vol. 9, no. 9, p. 1635, 2017. Available: 10.3390/su9091635 [Accessed 20 July 2020].

[2] J. Gutiérrez, J. García-Palomares, G. Romanillos and M. Salas-Olmedo, "The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona", Tourism Management, vol. 62, pp. 278-291, 2017. Available: 10.1016/j.tourman.2017.05.003 [Accessed 20 July 2020].

[3] D. Wang and J. Nicolau, "Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com", International Journal of Hospitality Management, vol. 62, pp. 120-131, 2017. Available: 10.1016/j.ijhm.2016.12.007 [Accessed 20 July 2020].

[4] P. Kalehbasti, L. Nikolenko and H. Rezaei, "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis", 2019. Available: https://arxiv.org/pdf/1907.12665.pdf. [Accessed 20 July 2020].

[5] E. Tang and K. Sangani, "Neighborhood and Price Prediction for San Francisco Airbnb Listings." [Accessed 20 July 2020].

[6] T. Cai, K. Han and H. Wu, "Melbourne Airbnb Price Prediction", 2020. [Accessed 21 July 2020].

[7] Y. Li, S. Wang, T. Yang, Q. Pan and J. Tang, "Price Recommendation on Vacation Rental Websites", Proceedings of the 2017 SIAM International Conference on Data Mining, pp. 399-407, 2017. Available: 10.1137/1.9781611974973.45 [Accessed 24 July 2020].

[8] "New York City Airbnb Open Data", Kaggle, 2019.

[9] P. Ye et al., "Customized Regression Model for Airbnb Dynamic Pricing", Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, 2018. Available: 10.1145/3219819.3219830 [Accessed 24 July 2020].

[10] H. Yu and J. Wu, "Real estate price prediction with regression and classification," CS229 (Machine Learning) Final Project Reports, 2016 [Accessed 22 July 2020].

[11] Y. Ma, Z. Zhang, A. Ihler, and B. Pan, "Estimating warehouse rental price using machine learning techniques.," International Journal of Computers, Communications Control, vol. 13, no. 2, 2018 [Accessed 22 July 2020]