Analysis of Airbnb Prices using Machine Learning Techniques

Jasleen Dhillon
California State University
Fresno, California

Nandana Priyanka Eluri
California State University
Fresno, California

Damanpreet Kaur
California State University
Fresno, California

Aafreen Chhipa
California State University
Fresno, California

Ashwin Gadupudi
California State University
Fresno, California

Rajeswari Cherupulli Eravi
California State University
Fresno, California

Matin Pirouz
California State University
Fresno, California

2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC) | 978-1-6654-1490-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/CCWC51732.2021.9376144

***Abstract*—Predicting a continuous variable with accuracy is not an easy task. In this paper, we initially perform data cleaning and data pre-processing. We, then, perform descriptive, prescriptive, and exploratory analysis to get a better understanding about the nature of the data. These analyses helped to understand the important attribute that needs to be taken towards the prediction of the price for our Airbnb listings. Even after performing data cleaning on the data set, there can be some outliers that are needed to be deeply examined, and therefore the outlier detection was performed on the dataset and flagged outliers were removed from the dataset. For predicting the prices, the models that were implemented are linear and Logistic regression and Random Forest. After implementing the above mentioned three approaches the best model was chosen based on the RMSE value of the model. In this paper, the least value of RMSE was for the Random Forest Method.**

***Index Terms*—Airbnb, listing, data, dataset, pre-processing, descriptive, predictive, exploratory, analysis, linear, logistic regression, random forest.**

I. Introduction

Airbnb stands for Air- bed, and breakfast. It is an online platform for posting a rental, main properties for tourism experiences. So essentially, Airbnb connects people looking to rent their houses with people who are looking for accommodations. The company does not own any kind of property, but they act as the third-party by receiving commissions from each booking made through them. Shaping a business as per the customer's dynamic is the most essential business tactic of the 21st century. The ones that benefit the most from this are the digital companies and one of those companies is Airbnb. People have vacant rooms, houses, hotels, or apartments at different locations that they want to lend to people who come for business trips or for tourism and simultaneously want their accommodations at low cost where a normal hotel would cost way more. The demand for the listings as per the city, most common property type preferred by the travelers, the price for the accommodation, duration of the stay, availability of the place, etc. are among many factors that help decide what finally drives a user to make the final reservation. Hosts have the option of adding price for any additional amenities they deem necessary. The host also needs to make sure that the prices are not too high to attract enough visitors. With the continuous rise in listings, coming up with the correctly priced listing to stay in the market is not just imperative but also competitive. Although, guests also struggle with an issue of their own - non-availability of accommodation. This can be because of many reasons such as seasonality, abrupt booking cancellations, or some change in preferences. We propose to analyze Airbnb's publicly available listing information for the past three years to try and validate our hypotheses to alleviate some of Airbnb's issues. It is very important for a host to get the price right for their properties and Airbnb site just provides general guidelines on how to price the properties. So it solely depends on the host to set their own prices for their apartment listings. There are a lot of factors that alter the daily prices around the base prices such as the availability in the area, number of people looking for a place, seasonality or which day of the week it is. We focus on knowing the graphs of the market of Airbnb and based on the past data we try to understand what the customers of Airbnb are looking for. Then, we co- relate all the results by making some important predictions for the reference of future customers. By implementing such methods, we can suggest customers plans and information regarding their tourism according to their past preferences and. Thus, this study not only helps the customers of Airbnb to meet their preferred accommodation but also helps Airbnb hosts by increasing the customers for their housing. Conclusively, this paper involves the analysis of the Airbnb dataset of different cities in the United States and visualizing the observations found in the analysis in order to improve the marketing and gain customer satisfaction by providing the required facilities. The paper is mainly focused on doing appropriate research about the Airbnb data derived from different and most popular cities (total 28 cities) across the US. The main objective of this paper is to make predictions about the prices for Airbnb housing, linking the prices with the locations and analyzing the minimum and maximum numbers of bookings per month.

978-0-7381-4394-1/21/$31.00 ©2021 IEEE

II. Related Work

Airbnb is a rental commercial business that has developed boundlessly in the recent couple of years. It has outpaced their opposition inns in giving momentary facilities to the visitors which makes it important to match the needs of the visitors to make them revisit the place [1]. With the increase in the Airbnb business, knowing how this will affect the travelers decides the future of innovative businesses. To look at the growth and development in the coming years, it is crucial to know the reasons which have led to change in society as a large number of platforms that provide accommodations were impacted [2]. It is crucial to analyze if the impact is based on location or the structure of the other businesses [3]. The fast pacing improvements have resulted in a great increase in the clash between Airbnb and various other businesses. There is a compulsion to improve and to diminish the expenses to keep the flow of visitors attracted to themselves. To match the steps with the business of Airbnb, it is important to have a proper understanding of the business model and numerous segments. The business models should be changed from time to time to maintain growth and development [3]. The Airbnb host should also be aware of what is expected from the property as compared to the hotels [2]. Various attributes of the listing affect the prices of Airbnb which makes it important to have a look at the relation between several attributes and how they differ while reflecting the price. The host thereby can decide the attributes which help the community grow more by keeping the check on the prices as well [4]. Analyzing the attitude of the hosts with the Airbnb property also plays a major role while renting [5]. Most of the people do not have information on how to price the property [4]. Therefore, pricing is one of the main factors affecting the accommodation system which makes it important to find the price determinant [6].

At the same time, it should be made sure that a particular platform like Airbnb needs to be regulated frequently. It can be seen that some of them are fully regulated whereas many do not establish any regulation, the decision about the financial state of a particular area needs to be taken after taking the regulations in account so that the visitors can take the advantage of this platform [7]. As we take various listing attributes into account, there is one really important factor which is the rating given to a particular place by the people who have already visited it once. Only ratings make the visitors trust that the area is worth living according to the recommendations of the community. It is important to compare the ratings of Airbnb with competitors of this business, in order to know the best results for a particular area [8]. Some of the main concerns regarding the sharing economy market are with the increase in the Airbnb business, and how much financial loss is faced by other businesses depending upon the change in the bottom line. At city level businesses, we distinguish Airbnb's effect by misusing noteworthy spatiotemporal variety in the examples of reception. We build up a nuanced gauge of Airbnb's material effect on lodging incomes. To segregate Airbnb's effect, we utilize lodging fragments that shoppers are less inclined to fill in for Airbnb remains as extra benchmark groups [9]. While dealing with Airbnb or any innovative business, it is important to include the sharing economy as it performs a major role in providing peer-to-peer accommodations. In other words, people can easily get the available accommodation in another person's property for a short time period, instead of paying for the hotel rooms. It is significant that the increase in the number of Airbnb's causes a decline in the number of people going to the hotel, making the hotel face the decline in their business [10]. As the number of Airbnb accommodations increased, the earnings of other businesses decreased [11].

The obligation has been put to strategize the pricing of the hotel accommodation in order to make the visitors stay in the hotels [12]. Another way of analyzing the growth of the business can be done by researching Airbnb from the double point of view of the visitors as well as the competitor businesses in order to know the motivation for the customers who choose to stay at the Airbnb instead of the hotels and also to find if the competitors are highly or rarely affected. The technology has also played a major role in the development of the Airbnb as this business provided the solution to various accommodation problems, making the customers behave as entrepreneurs, and making the people get familiar with the innovative idea of using the online services to book and decide where and what type of property a person needs in a particular area [13]. The selection of certain Airbnb properties is only done depending on the interaction between the host and the guest [14].

In order to note the difference between the prices of the hotel rooms before and after the Airbnb business started, we look at the factors affecting the prices. One of the main reasons is people choosing Airbnb over hotels, but for understanding the change better we need to look at what type of people now prefer to go to Airbnb, who usually went to hotels. It is also necessary to know if they permanently switched from going to the hotels to Airbnb, they use both Airbnb and hotels for accommodations, or they didn't switch to the use of Airbnb at all. It is helpful if we analyze that the reason for the use of Airbnb instead of hotels is just price or there are certain amenities that the visitors like more in the Airbnbs [15]. The Airbnb business needs to stay up to date to reduce the complaints and meet the needs of the visitors [16]. While handling all the complaints made by the hotels or other businesses, there is also concern shown regarding the prices of the properties in the residential areas. It is also a worry for the residents of that area how the people who come to visit the city affect the neighborhood as the increase in the number of visitors could make the prices for the residents go up and result in the place to be more expensive [17].

III. Method

A. Pre-Processing

We have the data, now the next step is to determine whether there exists a significant relationship between the variables in our data. Till this point in the paper, we have discovered that the price of a listing seems to be influenced

by multiple factors. After selecting the set of features, we try to develop a prediction model utilizing the regression analysis. This statistical technique is used for determining the relationship among a single dependent variable and one or more independent variables. In this project we will be talking about linear regression, logistic regression and random forest classifier followed by outlier detection. Let's talk about each of the machine learning techniques in detail.

B. Logistic-Regression

The goal of linear regression is to estimate the values for the model coefficients m_1, m_2, m_3...m_n and try to fit the training model with least root mean squared error and predict the values for the target variable. Similar predictions are done with logistic regression models but with one addition. The logistic regression model computes a weighted sum for the independent variables as the ones we took in linear regression but this time, the function used is a nonlinear function known as the logistic or sigmoid function and produces the dependent variable Y. The sigmoid/logistic function fits the following equation-

$$Y = \frac{1}{1 + e^{-1}} \tag{1}$$

Where,

$$Y = b_0 + (m_1 * x_1) + (m_2 * x_2) + (m_3 * x_3)...(m_n * x_n) \tag{2}$$

The output is interpreted as a binary value i.e. 1 or 0. The concept here is to predict the pricing of a listing right so that it will be rented out or not. Sometimes we are trying to see if a room is rented out or not to a guest based on the number of reviews that listing has, rating of the reviews, accuracy of the review scores, availability of the room, minimum and maximum number of nights a guest can spend there. In essence, it can be used as a target i.e. the model is predicting 1, when the probability of the variable is greater than 0.5 and when the probability is less than 0.5 then the classifier is predicting 0 [21].

C. Outlier Detection

Outliers in a dataset are those data points that deviate so much from the normal or average data points. These abnormal patterns in the dataset need to be examined carefully and to be removed from the dataset in order to increase the accuracy of the prediction model. The issue associated with this idea is that there is no labeled data available to use as a training data set for the anomaly detection, therefore an unsupervised machine learning algorithm is required for this approach. Hence, "One Class Support Vector Machine" has been implemented. One Class SVM is an unsupervised machine learning algorithm that can be used to classify the abnormal data points in a given dataset. One Class SVM method creates a decision boundary based on the given data and anything that lies outside of that boundary is classified as an outlier. The detection of the outliers is very necessary for the better prediction. In this project, prices of the listings for the Airbnb may contain these outliers and therefore can affect the overall performance of the predictive model. Hence, One Class SVM will help to detect the anomalies or outliers in the continuous variable price and improve the model accuracy for the further prediction analysis for the price.

IV. Data Visualization

First, the Airbnb dataset contains 305564 rows x 108 columns of data about various cities in the United States. Each city's dataset consists of the following datasets.

Listings - Contains listings for the specific city. It contains 108 attributes.

Reviews - Contains reviews given by the guests.

Calendar - Contains the details of the future bookings.

In order to start the data visualization, the data is cleaned as it consists of 108 variables, too many for the analysis, so the variables that are not used are removed. The dataset is further modified by removing the duplicated and non existing values. The variable like price is recognized as a factor data due to the "$" included in the data, the price variable is converted from object to numeric after removing the"$" symbol. Similarly, all the possible variables are converted to their respective formats for easy analysis. Also handled the correlations.

A. Visualizing Missing Values

To visualize the missing values, we have used a python library Missingio. This library helps us visualize the distribution of NaN values in the dataset. Missingio library helped us finding the correlation between the number of missing values in the different columns of the dataset as a heat map.

B. Exploratory Analysis

Number of new hosts and the first reviews on the Airbnb properties in the USA from 2008 to 2020. In "Fig. 1" the yellow line indicates the number of hosts joined and the green line indicates the first review of the listings. The oldest listing registered and available in the Airbnb USA was in March 2008. From the above graph we can see that the number of listings started to increase considerably from the year 2011. However, the number of hosts joining started to decrease from the year 2016.

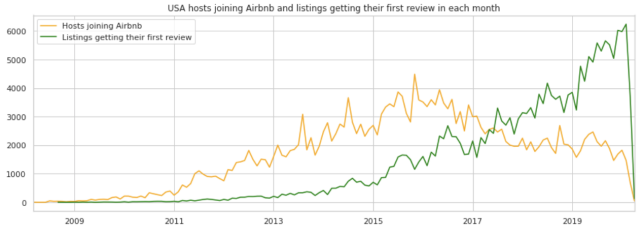

Fig. 1. Graph showing hosts joining the Airbnb listing and the first reviews

We can see that the number of hosts joining is mostly during the summer, to take advantage of attracting tourists during the summer holidays. The highest peak is by the end of 2015. As people have started to use the Airbnb properties, they have started giving reviews as well. From the above graph we can

see that the first reviews are increasing every year indicating that the number of people using the Airbnb properties are increasing every year. There is a drastic decrease in the number of reviews by the beginning of 2020, probably because of COVID-19. Due to this pandemic most of the tourist places have been closed and people in the USA have been asked to shelter in place. "Fig. 2" shows the time series decomposition of the number of hosts joining the Airbnb USA listings.

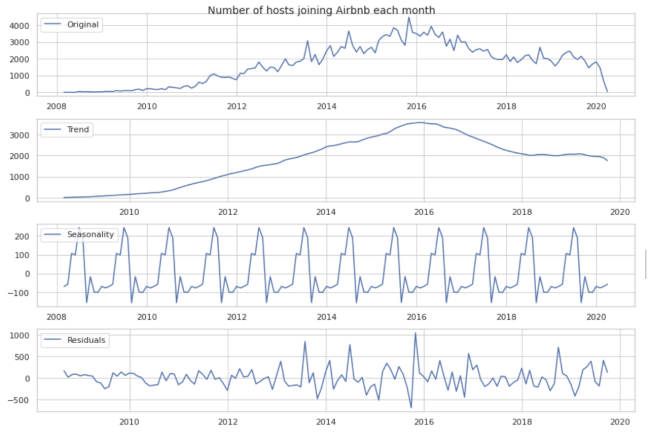

Fig. 2. Hosts joining in each month in the Airbnb USA

This logarithmic data transformation is used to smooth out the distributions. Now, the correlations between the variables start to emerge. The graph in "Fig. 3" displays the number of

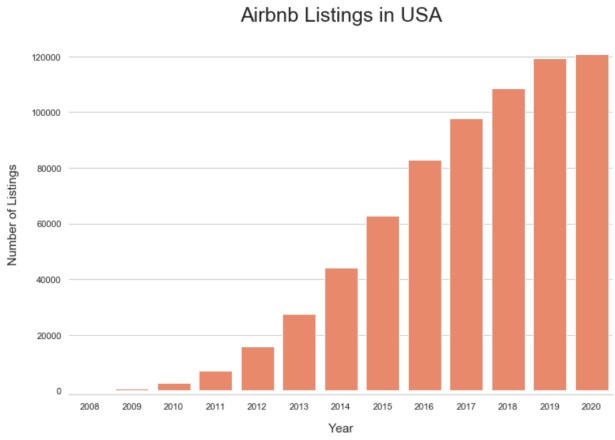

Fig. 3. Airbnb Listings in USA from some top cities in different years

Airbnb listings in some top cities in the USA from the year 2008 - 2020. From the below bar graph, we can see that the number of Airbnb properties has increased by a large number. In 2008 we could not see anything because there could be just a few 100s of properties available. The number of properties has increased at a similar rate.

"Fig. 4" gives us an overview of the property types listed in Airbnb listings in the USA. From the below graph we can see that the number of apartments/entire house is the highest listed property which is more than 80000. The least

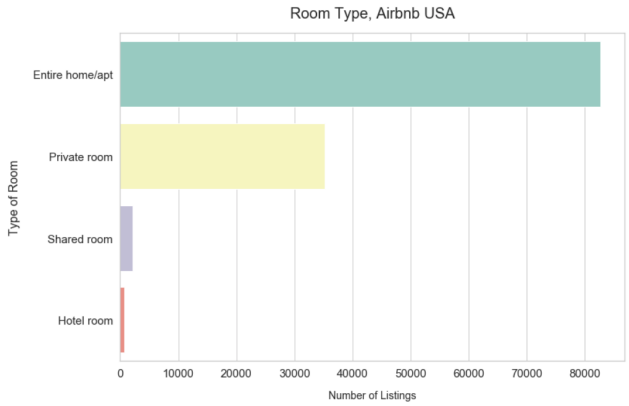

Fig. 4. The available property types in the Airbnb listings in the USA

number of listings available in Airbnb is the hotels. Hotels and shared rooms are comparatively lesser than other property types namely, private rooms and entire home/apartments.

C. Descriptive Analysis

From "Fig. 5", we can see about 68% of the Airbnb listings are of type entire home/apartment (i.e., own the entire property) and 29% are private rooms (i.e. you stay with other people in the property but own your own room and bathroom). About 41% of available properties are apartments. Other properties are houses and uncommon property types. Nearly 170000 rooms are of listing Entire home/apartment and 71000 rooms are private rooms. Whereas shared rooms and hostels are very less in number because they are available to book outside or directly without using Airbnb.

We can clearly see from "Fig. 6", that the average price for Airbnb listings in the United States has increased drastically over time from the past 10 years. Particularly, high property prices are increased resulting in large increase in the mean price when compared to median. For example, the mean price in 2009 was $123 whereas in 2019 it was $237.67.

"Fig. 8" shows the most common time period, the Airbnb listings had their first review in two - three years. This means most of the listings are active for at least a couple of years. Few listings are active for more than four years.

The most common category is in 1+ years for the time since a listing received its last review. A lot of listings have been reviewed relatively last year. But still there are a lot of listings which have not had a review for more than a year. These majority listings are termed as "inactive" listings, because although they are technically on the site, but do not have openings on their calendars open and are not available to book.

V. Preliminary Results

By the descriptive analysis, we have established that there are a lot of variables in our dataset which can help in prediction of the prices of the Airbnb properties listed. Initially, a linear assumption is made for the same purpose. Based on this assumption that the data points are linearly distributed following

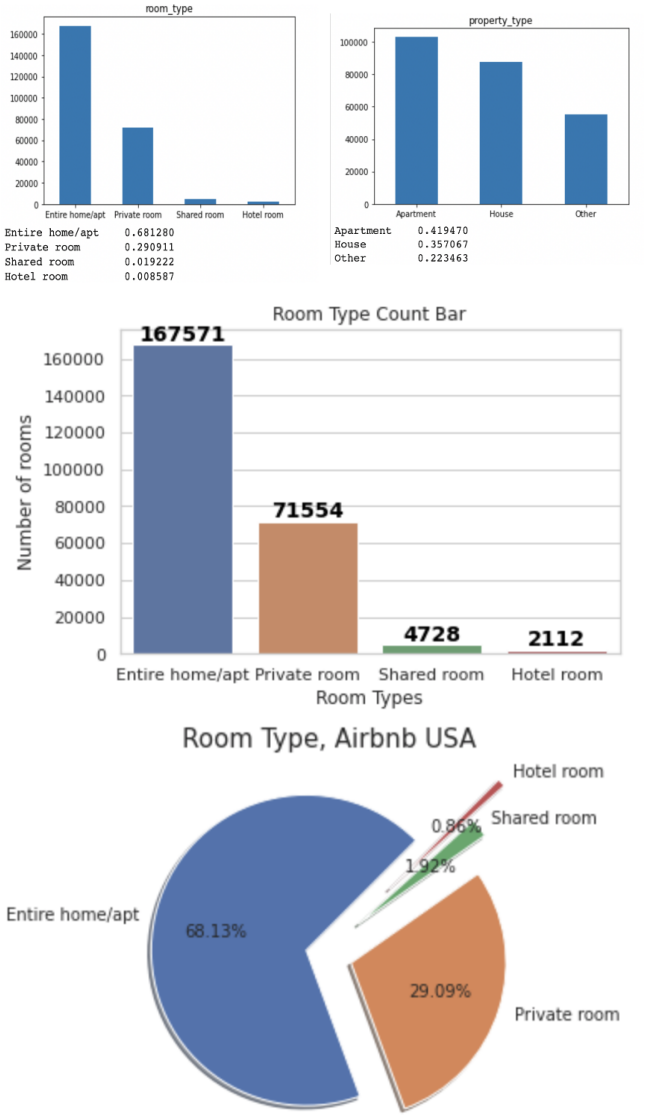

Fig. 5. Most common property and room type in the Airbnb listings

Fig. 6. Different patterns of change of price over time

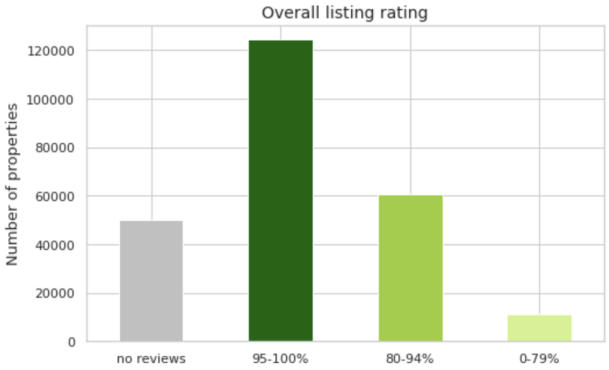

Fig. 7. Overall Rating for Airbnb listings

plots are made for different variables with respect to price. The variables will be in the order - city, latitude, longitude, accommodations, bathrooms, bedrooms, number_of_reviews and review_scores_rating fitted along with price.

As per the linear assumptions and correlation matrix values for columns such as latitude, longitude, number_of_reviews and review_scores_rating we conclude that they are not making much impact on price and we drop them as shown in the "Fig. 9".

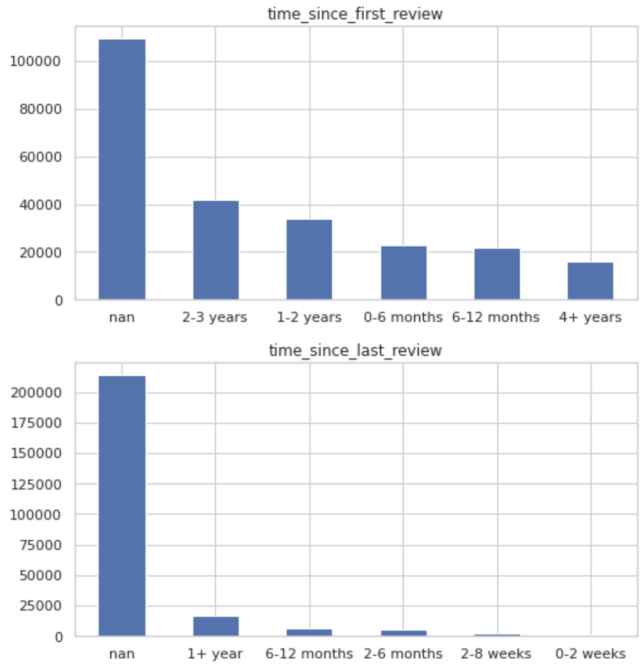

Fig. 8. Listings on site and have been reviewed

After implementing the linear regression model with the variables left, we get the following outcomes –

Size of the training model – 80%

Size of the testing model – 20%

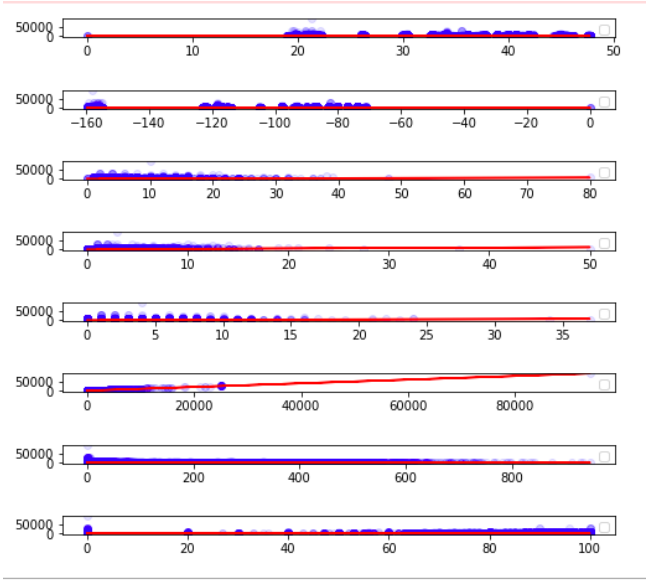

Fig. 9. Scatter plots for city, latitude, longitude, accommodates, bathrooms, bedrooms, number_of_reviews and review_scores_rating fitted along with price

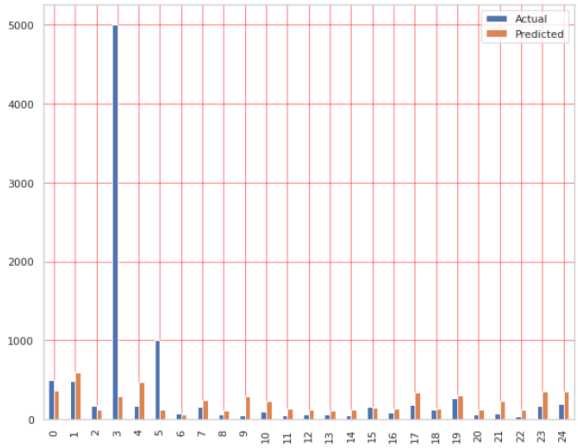

Fig. 10. Actual prices versus prices predicted by the linear regression model

VI. Results and Analysis

A. Linear Regression

"Fig. 10" shows the difference in the actual prices and the prices predicted by the linear regression model. After looking at the graph, we are able to analyze how the actual price varies from the predicted prices by which we are able to calculate the mean squared error to check how accurate the model is for predicting the prices based on certain features.

B. Logistic Regression

For the logistic regression, instead of plotting a linear relation between the independent and dependent variables we divide the data set into the train and test data to get the logistic relation. In our project, we divide the training and testing data as 80% and 20% respectively. In this model, we first train some part of the data for which the prices are already given. Then, we get a logistic equation representing the probability of the price. After getting the probability of the prices using the trained data, we predict the prices for the testing data by using the logistic equation which was found using the training data. The accuracy of the logistic regression classifier on the test set is 0.78 as shown in table 1.

C. Random Forest Classifier

In this model, we used the minimum_nights, maximum_nights, availability_365, number_of_reviews, review_scores_rating, review_scores_accuracy, accommodates for the prediction of the price. This model estimated the prices after fitting various decision tree classifiers on different subtests of the data set. The final answer is provided after averaging the predictions in order to improve precision and power over-fitting. We were able to get the accuracy of 87% using the random forest model as shown in table 2.

	Actual	***Predicted***
94490	137.0	61.535406
294002	165.0	128.099949
107326	150.0	201.378376
224551	80.0	77.263397
259121	165.0	74.868581
81871	59.0	159.812030
124058	90.0	164.129657
297145	100.0	60.104383
1473	75.0	222.226527
73781	725.0	431.488294
244774	92.0	46.532529
135766	125.0	226.569296
282306	95.0	-18.770521
92784	130.0	405.325264
122833	125.0	58.650028
190226	60.0	109.719889
99485	113.0	94.887807
130005	150.0	399.465329
188785	135.0	91.449145
210094	75.0	83.157640

Fig. 11. Comparing the predicted values for the price column in the dataset

TABLE I
Classification Report for Logistic Regression

	Precision	***Recall***	***F1 Score***	***Support***
0.0	0.78	1.00	0.88	38461
1.0	0.00	0.00	0.00	10732
Accuracy			0.78	49193
Macro Avg.	0.39	0.50	0.44	49193
Weighted Avg.	0.61	0.78	0.69	49193

Fig. 12. Logistic Regression results

Fig. 13. Random forest regression results

VII. Conclusion

Our project is mainly focused on doing proper research using the Airbnb data from different cities across the US. The main goal of this project is to make predictions of the prices for Airbnb using different machine learning algorithms. Our motivation is also to find out which Machine Learning algorithm works best for this dataset in predicting the prices with maximum accuracy. In our research, we have found out that Random forest model works best with this kind of data. The project will be providing the host with the best information according to their particular needs such as location, ratings, and price. The accuracy of the prediction models can be further improved using the reviews and the listing summary attributes.

TABLE II
Classification Report for Random Forest

	Precision	*Recall*	*F1 Score*	*Support*
0.0	0.89	0.95	0.92	38461
1.0	0.75	0.58	0.66	10732
Accuracy			0.87	49193
Macro Avg.	0.82	0.76	0.79	49193
Weighted Avg.	0.86	0.87	0.86	49193

References

[1] Mao, Zhenxing, and Jiaylng Lyu, "Why travelers use Airbnb again?" Internationl Journal of Contemporary Hospitality Management (2017).

[2] Li, Yang, et al. "Price Recommendation on Vacation Rental Websites." Proceedings of the 2017 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2017.

[3] Moon, Hyoungeun, et a1. "Peer-to-peer interactions: Perspectives of Airbnb guests and hosts." International Journal of Hospitality Management 77 (2019): 405-414.

[4] Sheppard, Stephen, and Andrew Udell. "Do Airbnb properties affect house prices." Williams College Department of Economics Working Papers 3.1 (2016): 43.

[5] Dogru, Tarik, Makarand Mody, and Courtney Suess. "Adding evidence to the debate: Quantifying Airbnb's disruptive impact on ten key hotel markets." Tourism Management 72 (2019): 27-38.

[6] Wang, Dan, and Juan L. Nicolau. "Price determinants of sharing economy-based accommodation rental: A study of listings from 33 cities on Airbnb.com." International Journal of Hospitality Management 62 (2017): 120-131.

[7] Oskam, Jeroen, and Albert Boswijk. " Airbnb: the future of networked hospitality businesses." Journal of Tourism Futures (2016).

[8] Quattrone, Giovanni, et a1. " Who benefits from the "Sharing" economy of Airbnb?" Proceedings of the 25th international conference on the world wide web. 2016.

[9] Zervas, Georgios, Davide Proserpio, and John Byers. "A first look at online reputation on Airbnb, where every stay is above average." Where Every Stay is Above Average (January 28, 2015) (2015).

[10] Varma, Arup, et al."Airbnb: Exciting innovation or passing fad?" Tourism Management Perspectives 20 (2016): 228-237

[11] Guttentag, Daniel A., and Stephen LJ Smith. "Assessing Airbnb as a disruptive innovation relative to hotels: Substitution and comparative performance expectations." International Journal of Hospitality Management 64 (2017): 1-10.

[12] Dogru, Tarik, et a1. "Does Airbnb have a homogenous impact? Examining Airbnb's effect on hotels with different organizational structures." International Journal of Hospitality Management 86 (2020): 102451

[13] Goree, Katherine. "Battle of the beds: the economic impact of Airbnb on the hotel industry in Chicago and San Francisco"

[14] Moon, Hyoungeun, Wei Wei, and Li Miao. "Complaints and resolutions in a peer-to-peer business model." International Journal of Hospitality Management 81 (2019): 239-248.

[15] Bashir, Makhmoor, and Rajesh Verma. " Airbnb disruptive business model innovation: Assessing the impact on the hotel industry." International Journal of Applied Business and Economic Research 14.4 (2016): 2595-2604.

[16] Chua, Evelyn L., Jason L. Chiu, and Nelson C. Bool. "Sharing Economy: An Analysis of Airbnb Business Model and the Factors that Influence Consumer Adoption." Review of Integrative Business and Economics Research 8 (2019): 19.

[17] Neeser, Dâvid, Martin Peitz, and Jan Stuhler. "Does Airbnb hurt hotel business: Evidence from the Nordic countries." Universidad Carlos III de Madrid (2015): 1-26.

[18] Gibbs, Chris, et a1. "Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings." Journal of Travel Tourism Marketing 35.1 (2018): 46-56.

[19] Chua, Evelyn L., Jason L. Chiu, and Nelson C. Bool. "Sharing Economy: An Analysis of Airbnb Business Model and the Factors that Influence Consumer Adoption." Review of Integrative Business and Economics Research 8 (2019): 19.

[20] Roma, Paolo, Umberto Panniello, and Giovanna Lo Nigro. "Sharing economy and incumbents' pricing strategy: The impact of Airbnb on the hospitality industry." International Journal of Production Economics 214 (2019): 17-29.

[21] Tayeb, S., Pirouz, M., Sun, J., Hall, K., Chang, A., Li, J., ... & Latifi, S. (2017, December). Toward predicting medical conditions using k-nearest neighbors. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 3897-3903). IEEE.