How to Become a Top Earner on Airbnb ?

Anushka Pandey
PES1UG20CS072
Computer Science & Engineering PES
University Bangalore, India
anushkapandey1105@gmail.com

Devika S
PES1UG20CS123
Computer Science & Engineering PES
University Bangalore, India
reachdevikas@gmail.com

***Abstract*— The primary goal of this project is to analyze the Airbnb data statistics of a popular seaport city – Seattle, in the US. It uses exploratory data analysis and data visualization to bring out the patterns within the numerous attributes of the data set. This project also comprises a model which can predict the rent in a particular neighborhood and also the availability of rooms.

Analysis of such a data set is necessary so as to gain insight of the current demand of the tourists, so that we can implement such a setup in our country as well.

I. Introduction

Tourism is one of the most important sources of income for all the nations. It is observed that few nations like Maldives, Bali,Singapore,Nepal etc. rely only on tourism for their economic growth. With tourism comes the responsibility of making the guests comfortable and giving them high standards of facilities, with the rooms and hotels being the most important aspect of their trip. Airbnb Inc. is an American company which is famous for providing lodgings for vacation stays and other tourist activities.

It provides rental options for different customer segments. Based on their budget, people can rent an entire house or just a room by just a click of a button. Apart from that users can also enter their reviews about their stay at the place and other facilities like cleanliness, location, communication etc. A major question answered in this project is – ‘ how is Airbnb affecting the neighborhoods ? ’ . By the help of our project, we can predict ‘when are the prices high and low’ , ‘which neighborhoods are best preferred by the tourists’, ‘relation between average tourist volumes and months of the year’ etc. We have also shown relationships between different attributes of the dataset by the aid of pairplots, histograms, heatmaps, correlation matrix etc.

II. Review of Literature

A. *The rise of exponential organization*

Analyzing [3], we realize that the digitalization of today’s society has made it possible for organizations to reach beyond the traditional markets. Exponential markets grow 10 times faster than their traditional counterparts and perform better and faster. Ismail et al.(2014)[4] describe the success of these organizations extensively, notable among these are - they have a compelling higher purpose and they dare to experiment, manage algorithms and have empowered autonomous workers. Airbnb, Uber, Gitgap are some of the renowned “exponential organizations”.

B. *The rise of Sharing or Collaborative Economy*

Parties like Airbnb and Uber identify themselves as sharing economies wherein the consumers grant each other short term access to underutilized physical assets like rooms and apartments in case of Airbnb and cars in case of Uber. Transfer of ownership is not done in such scenarios. This way it benefits both the temporary customer and the owner. The advent of internet has also made this economy flourish to large extents.

C. *Spike in demand for rental homes*

The surge in tourism has led to an increased demand for rental homes and accommodations. Thus, the market for rental homes has widened. Airbnb has captured a huge portion of this market. That is why, this paper aims at finding the methods to become a top earner on the Airbnb platform and maximize revenue and help customers make right decisions when choosing their rooms/apartments. This will also enable the hosts to make smart business decisions.

D. *Using Multiple Linear Regression*

From [5], we see that Multiple Linear Regression has been used to predict the price of houses. Significant factors/features that influence the response variable are selected and then the model is built to predict the price. Similarly, we plan on predicting the price of rental rooms after selecting significant attributes like neighborhood, rating, type of room, etc.

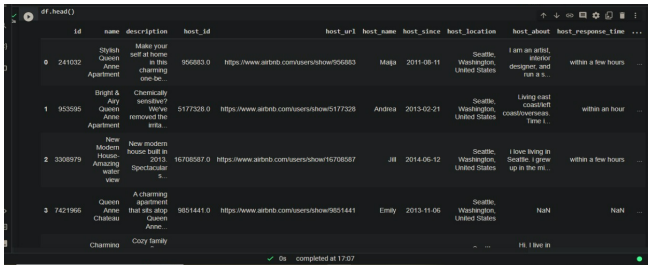
III. Dataset

The chosen dataset, **‘ Seattle Airbnb Open Data ’** has been procured from [1].
https://www.kaggle.com/datasets/airbnb/seattle
This dataset has three csv files – calendar.csv, listings.csv and reviews.csv. We see that there are 92 variables and 3818 observations for listing Some of the data was collected by scraping Airbnb websites. Airbnb is a $75 Billion online marketplace for renting out homes/villas/apartments/private rooms. The website charges a commission (3 to 20 percent) for every booking. We can infer quite some details not only about Airbnb’s business and their hosts, but also about Seattle. A lot of insights can be drawn from this dataset that can help improve the host’s business. This will help us answer questions like which neighborhood is popular among tourists and identify neighborhoods to invest in(to maximize the estimated revenue). We also believe we can get a rough idea about the relationship between Airbnb as a listing service and the hosts. We will be able to predict what factors(like cleanliness, amenities provided, surroundings) do customers consider most important while looking for a

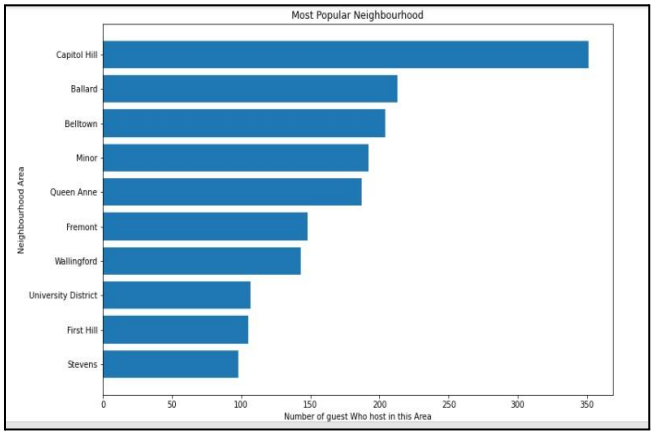
place to rent. This can be used to make informed business decisions.

IV. Initial Insights

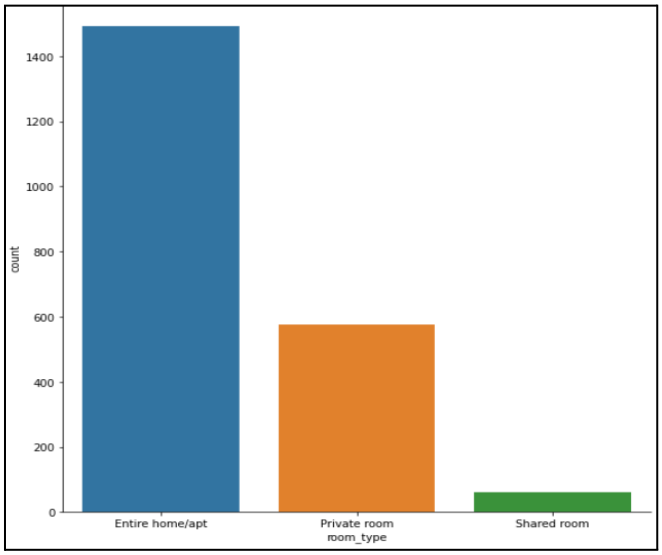
After observing the dataset and having cleaned the dataset, we now perform Exploratory Data Analysis(EDA), with the intention of gaining insights from this process. Null values have been replaced by appropriate metrics.



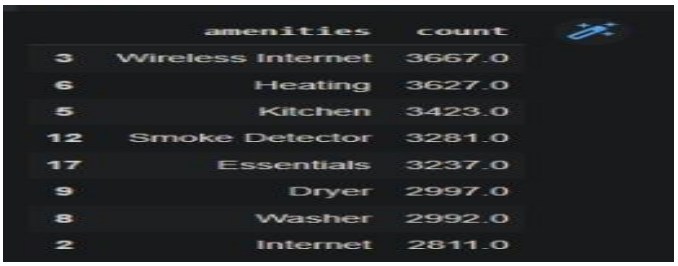
A glimpse of our dataset after the required cleaning and preprocessing



Using the library Matplotlib, we have generated the above plot, which illustrates the popularity of the various neighborhoods in Seattle. Capitol Hill appears to be the most popular neighborhood.

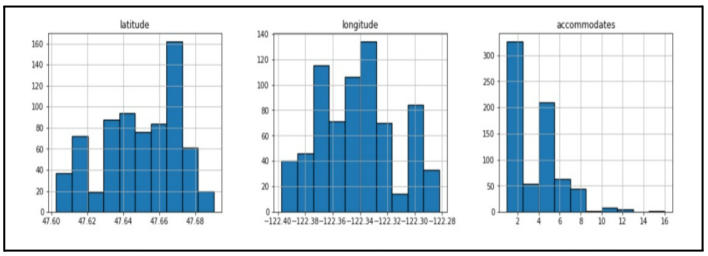


The above categorical plot was created using the Seaborn function catplot. We infer from this plot that most of the listings are of entire homes/apartments, followed by private rooms.

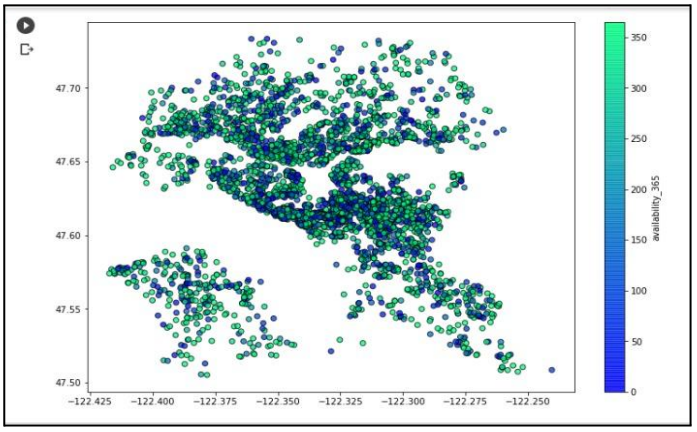


	amenities	count
3	Wireless Internet	3667.0
6	Heating	3627.0
5	Kitchen	3423.0
12	Smoke Detector	3281.0
17	Essentials	3237.0
9	Dryer	2997.0
8	Washer	2992.0
2	Internet	2811.0

The above table shows the number of amenities provided by the hosts all over Seattle.



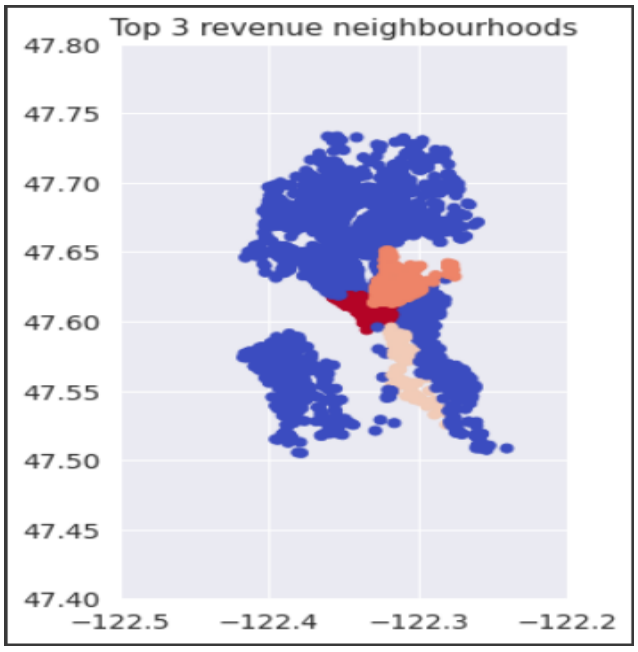
Histogram of a few attributes like latitude and longitude



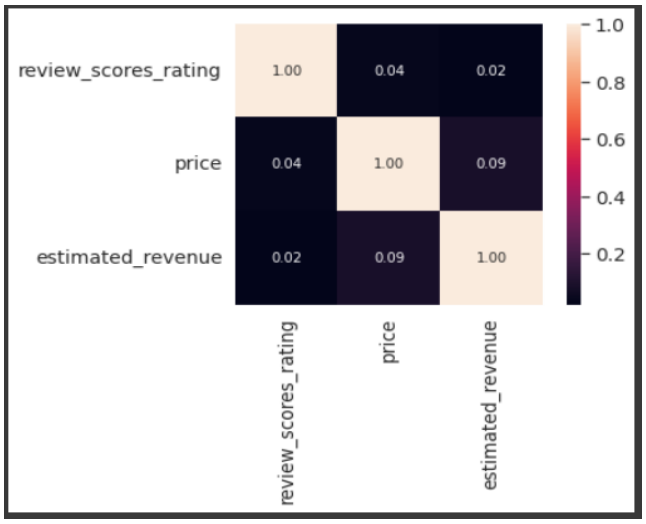
The above scatter plot is plotted between latitude and longitude, taking availability as the hue. We can see distinct clusters. It shows the availability of rooms around the year in different neighborhoods, based on their locations.

neighbourhood_group_cleansed	estimated_revenue
Downtown	7247.666038
Capitol Hill	7064.079365
Beacon Hill	6606.983051
Ballard	6078.726087
Queen Anne	6038.752542
Central Area	4636.371274
Other neighborhoods	4409.843829
Cascade	4075.134831
Seward Park	4063.500000
Rainier Valley	3827.345912
Delridge	3641.189873
Magnolia	3587.819672
West Seattle	3370.783251
Northgate	2962.362500
Lake City	2476.432836
Interbay	2105.727273
University District	1558.557377

The above table shows the overall estimated revenue of listings in each neighborhood.



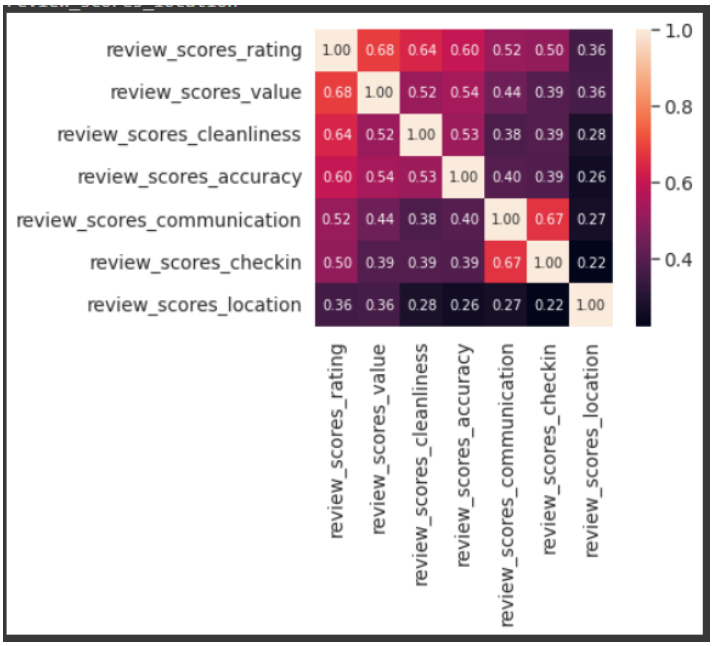
So the inference that can be made from the above table and scatter plot is that the neighborhoods of Downtown, Capitol Hill and Beacon Hill have the highest revenue.



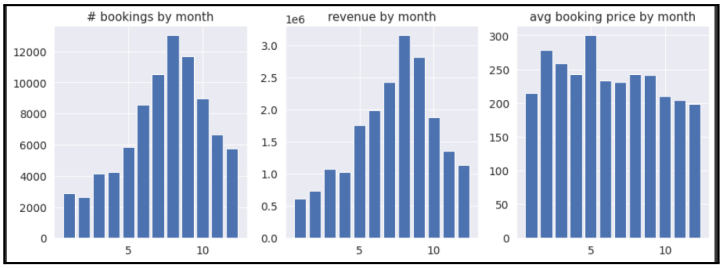
From the above correlation matrix, we can infer that -

1. The overall rating (described by the attribute review_scores_rating) has a very small positive correlation with the price. This means that overall rating does not change on changing the price set by the host. This implies that an inexpensive room can have a high overall rating and an expensive room can have a low rating. An expensive room can also have a high rating and an inexpensive room can have a low rating.

2. The overall rating (described by the attribute review_scores_rating) has a very small positive correlation of 0.02 with the estimated revenue. This implies that having a high overall rating does not guarantee a high estimated revenue. The overall rating column had several missing values. This could possibly be explained by the fact that most guests prefer not to rate the rooms and services.



1. The above correlation matrix has been used to analyze the correlation between overall ratings and other rating parameters like cleanliness, location, communication and others. We can see that overall ratings have a strong positive correlation with cleanliness and accuracy.



The above plots describe the number of bookings by month, revenue per month and the average booking price per month. The number of bookings per month and revenue per month are slightly skewed to the left. From these plots, we can infer that July, August and September are the best months to maximize revenue as most of the rooms are booked during these months.

V. Problem Statement and Approach

The problem statement is “How to become a top earner on the platform Airbnb?” We intend to compare top performer's acts (the ones that have more bookings and great listing's rating) to the low performers and figure out the methods the latter can opt to improve their performance. This can also be used by hosts to increase services provided by them to create “premium” features. We also seek answers to questions like which the most popular neighborhood is, whether the ratings of a listing are affected by the price and see if we can increase the estimated revenue.

VI. Methodology

After preprocessing the data and performing exploratory data analysis, we now have a good understanding of the dataset. A recommendation system has been built. Several machine learning models have been implemented to predict the price.

The training - testing split was taken as 80%-20%. EXtraTreeRegressor has been used to select the features that are the most important to predict the price. The various models that we have implemented and their details have been listed below.

A. Recommendation System

The recommendation system built by us can prove useful to both hosts and customers. The input to the recommendation system is the listing id and the number of similar listings desired (k). The system outputs k listings which are similar to the input listing along with its location and description. This model uses the description of the listings to calculate similarity scores between listings. Hence it is a content-based recommendation system. TfidfVectorizer has been used. Stop words have been neglected. Then on the basis of the term frequencies of the other words our model is trained.

A new host can figure out the services or prominent details from the description of the k listings that are most similar to the given listing (entered as the most popular one) and can thus choose appropriate locations and services to set up their apartment.

B. Multiple Linear Regression

Multiple linear regression is a linear regression model that estimates the linear relationship between a dependent variable and two or more independent variables. The features used are bedrooms, accommodates, bathrooms, beds and square feet. These features were selected on the basis of their correlation with price, which was visualized using a correlation matrix.

The idea is that the host can enter the number of bedrooms, beds, bathrooms, etc and decide how to price their listing. This will help them price the room/apartment appropriately. Hosts can also decide how many bedrooms, bathrooms to have in order to set a certain price and thereby have a certain earning. The metric used is the R-squared value.

C. Support Vector Regressor

Support Vector Regressor is a powerful supervised learning algorithm that can be used for regression. We have used the radial basis function (rbf) as the kernel function.

The five attributes used to predict price are selected on the basis of the output of the Extra Tree Regressor, which includes, bedrooms, accommodates, bathrooms, beds and latitude.

D. XGBoost

Gradient boosting is a supervised learning algorithm that can be used to accurately predict a target variable by combining an ensemble of estimates from a set of simpler and weaker models.

From the results of the ExtraTreeRegressor, it is evident that bedrooms, accommodates, bathrooms, beds and latitude are the important features. Hence, we have used these features to predict price. We train the model using eXtreme Gradient Boosting on the above listed features to predict price.

We have trained this model with n_estimators set to 20 and the learning rate set to 0.1. The root mean squared error is the calculated metric.

E. RANSAC Regressor

Regression models using least squares estimation are very sensitive to outliers. So to overcome this problem we can use the RANSAC algorithm .It is an iterative algorithm for the robust estimation of parameters from a subset of inliers from the dataset.

The main goal of the model is to predict the price on the basis of the most important features of a listing like bedrooms, bathrooms, beds, accommodates and location. These features were determined using the Extra Tree Regressor algorithm. Finally the mean squared error is calculated to determine the accuracy of the model.

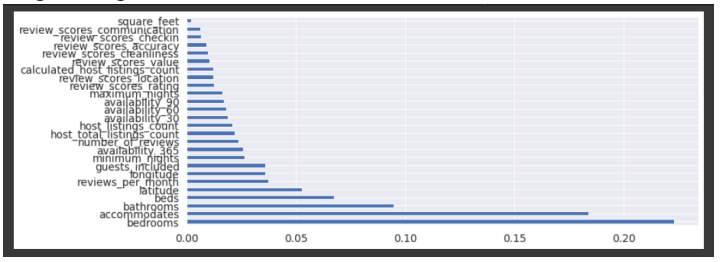
F. Random Forest Regressor

Random Forest Regressor is a supervised machine learning algorithm. It is a bagging ensemble technique that combines the predictions from multiple trees to make a more accurate prediction than the individual trees.

The five attributes used to predict price are bedrooms, accommodates, beds, bathrooms and latitude. 20 decision trees have been used.

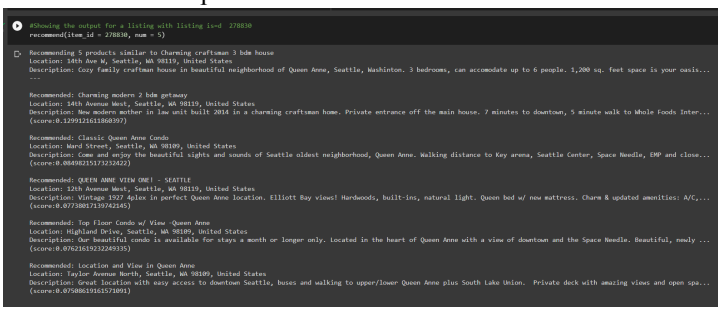
VII. Results

Results of the Extra Tree Regressor have been shown below. Only the numeric/non-categorical attributes have been used to choose the most important features to predict price.

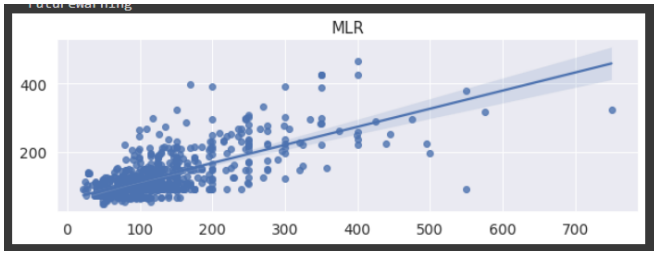


A. Recommendation System

The results of the recommendation system for listing id = 278830 has been shown below. 5 listings similar to the listing input have been printed, along with their location and description.



B. *Multiple Linear Regression*



The scatter plot of the MLR to predict price has been shown above. There seems to exist some amount of linearity between the five chosen features and price. The RMSE score of the implemented MLR model has also been calculated on the validation set and has been presented below. The RMSE value is 60.33

```
[169] print("R^2 value using score fn: %.3f" % lm.score(X_test,y_test))
      print("Root Mean Squared Error : %0.3f" % np.sqrt(mean_squared_error(y_test,y_pred)))

      R^2 value using score fn: 0.482
      Root Mean Squared Error : 60.333
```

C. *Support Vector Regressor*

The SVR model uses the rbf function as the kernel function. The RMSE score of the implemented SVR model has also been calculated and has been presented below. The RMSE value turns out to be pretty high, 156.45.

```
[205] rmse = np.sqrt(mean_squared_error(y_test, y_pred))
      print("RMSE: %f" % (rmse))

      RMSE: 159.453780
```

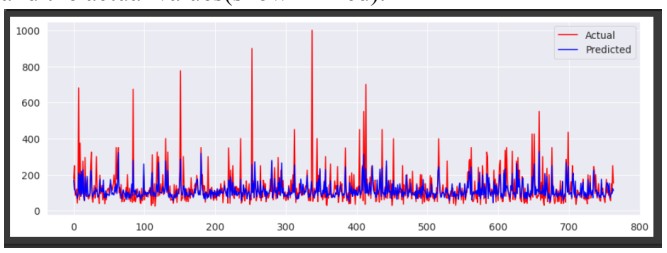
D. *XGBoost*

The RMSE score of eXtreme Gradient Boosting has been shown below. The RMSE score is better than the SVR. The RMSE score is 73.051

```
[180] rmse = np.sqrt(mean_squared_error(y_test, preds))
      print("RMSE: %f" % (rmse))

      RMSE: 73.050683
```

The below plot shows the difference in the predictions made by the XGBoost model(shown in blue) and the actual values(shown in red).



E. *RANSAC Regressor*

The RMSE value of the RANSAC regressor turns out to be 91.346

```
[193] rmse = np.sqrt(mean_squared_error(y_test, y_pred))
      print("RMSE: %f" % (rmse))

      RMSE: 91.345684
```

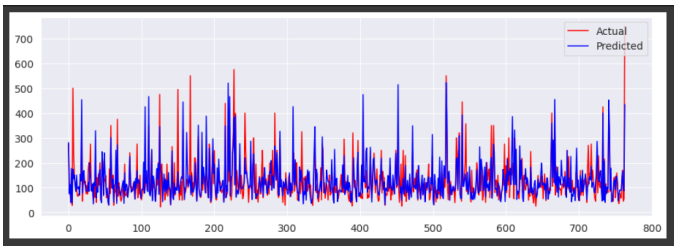
F. *Random Forest Regressor*

The RMSE value of the RANSAC regressor turns out to be 58.838, which is the best out of all the models tested so far. 20 decision trees have been used to build this model.

```
[174] y_pred = regressor.predict(X_test)
      print("Root Mean Squared Error : %0.3f" % np.sqrt(mean_squared_error(y_test,y_pred)))

      Root Mean Squared Error : 58.838
```

The below plot shows the difference in the predictions made by the Random Forest Regressor(shown in blue) and the actual values(shown in red). From visual inspection of this plot, it is evident that this model performs better than the XGBoost model.



Now, we are comparing the RMSE metric of the various models tested to predict price. These have been listed in the below table. We can observe that Random Forest Regressor performed the best, with a RMSE score of 58.838, while SVR performed the worst, with a RMSE score of 159.454

S.No	MODEL	RMSE
1.	Random Forest	58.838
2.	MLR	60.333
3.	XGBoost	73.051
4.	RANSAC	91.346
5.	SVR	159.454

VIII. Conclusion

This report encapsulates all the tasks which we have done while analyzing our dataset so as to find out ways to tackle our problem statement that is - 'How to Become a Top Earner on Airbnb?'

We started with the EDA of our dataset and found out some intricate details and dependencies among the attributes in our dataset. We then performed a literature survey and analyzed the available literature on economies and demand for rental homes in modern days. These activities helped us in selecting our models for the training of our dataset. The models in turn predict the price that the host must set based on the number of bedrooms, beds, bathrooms and accommodates and based on the latitude.

As we can infer from the above table, the Random Forest regressor gives the least RMSE value (=58.838) and thus it is considered to be the best algorithm to train our model for price prediction. In the plot of Random Forest Regressor , the predicted values also seem to coincide with the actual value much better than the ones in the plot of the XGBoost model.

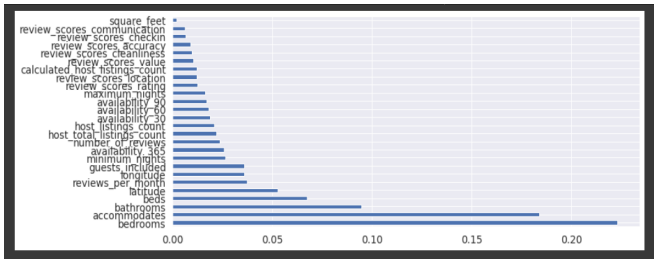
We have answered questions like which months have the most number of bookings, which amenities are the most common ones, which neighbourhoods generate the most revenue and which ones are the most popular. We have also implemented models to We have also implemented a recommendation system which can help both tourists and hosts.

IX. Peer Review

The suggestions given to our team during the peer review have been listed below:

1. Use Extra Tree Regressor for feature selection:

This has been implemented by us to choose the most important features to predict price. Previously, we had used the correlation between the various attributes and price to choose the important features. The plot obtained has been shown below. The 5 most important features according to the Extra Tree Regressor are bedroom, accommodates, bathrooms, beds and latitude.

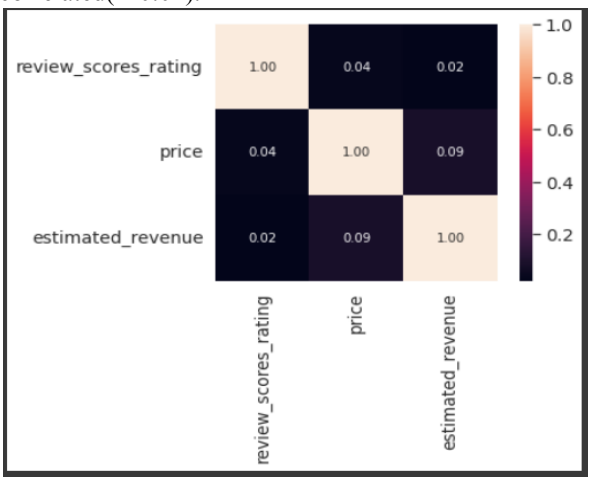


2. Time Series analysis

We have not implemented this suggestion as our dataset includes date and price attributes of several listings. Hence, it would not give us insights about how price changes with time as the price given is of several listings.

3. Correlation plot

The correlation plot between rating and price has been plotted below. It is evident that price and review_scores_rating are very weakly positively correlated(r=0.04).



X. Contributions of Team Member

Both of us performed exploratory data analysis (EDA) and then we discussed our conclusions from that. Further for the next stage, each person worked on different models for predicting the price. The exact details of the work distribution are given below -

Devika S - MLR, SVR, XGBoost
Anushka Pandey - Recommendation System, RANSAC Regressor, Random Forest

Both of the team members were involved in report writing. Most of the discussions were done on a google meet.

Acknowledgment

We would like to express our profound gratitude to Dr. Gowri Srinivasa and the entire Data Analytics team, for encouraging and providing us with this opportunity to get hands-on experience in the field, and guiding us along the way. We would also like to thank the Computer Science and Engineering department at PES University, for always inspiring us to conduct frequent research and inculcating a problem-solving discipline in us.

References

[1] https://www.kaggle.com/datasets/airbnb/seattle

[2] https://towardsdatascience.com

[3] https://www.researchgate.net/publication/298305479_Airbnb_The_future_of_networked_hospitality_businesses

[4] Ismail, S., Malone, M., van Geest, Y. and Diamandis, P. (2014), Exponential Organizations: Why New Organizations are Ten Times Better, Faster and Cheaper than Yours (and What to Do About It), Diversion Books, New York, NY

[5] https://doi.org/10.1155/2021/7678931

[6] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RANSACRegressor.html

[7] https://www.geeksforgeeks.org/xgboost/