Superhost Qualities and Future Price Prediction for Airbnb

Divyank Rahoria
A53284395

Tasks:

1: Dataset Identification and exploratory data analysis:

The dataset chosen for this assignment is from noncommercial and independent website insideairbnb.com. This website provides a set of tools and data that allows to explore how Airbnb is really being used in cities around the world. These datasets are collected by Airbnb website's listing but not by Airbnb. The website keeps updating this dataset as Airbnb listing s are growing by each day. The dataset I used for the assignment has updated listings till September 14th, 2019. By analyzing publicly available information about a city's Airbnb's listings, Inside Airbnb provides filters and key metrics so we can see how Airbnb is being used to compete with the residential housing market. Inside Airbnb provides the data for the research purposes open to all. Following figure shows the Airbnb property spread in Paris provided by Inside Airbnb.

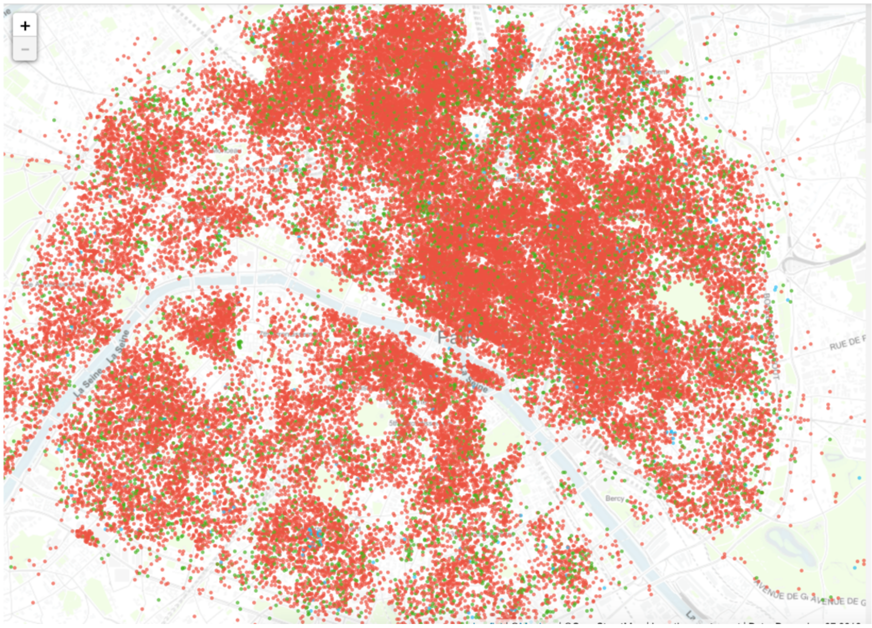

Fig.1: Airbnb listings in Paris (PC. Insideairbnb.com)

As the dataset is very latest, I didn't see this dataset used in some data science application, so I chose this for my assignment. This dataset has 64790 Airbnb listings in Paris that is above the requirements for the assignment.

Basic Statistics and Properties:

Inside Airbnb collects all the information provided publicly on the Airbnb website, it was able to provide us with unique features regarding each Airbnb listing. We can quickly see that this dataset included a mixture of qualitative, quantitative, and repeated information and unnecessary information. The qualitative data included the summary and descriptions, ratings of users the host provided. Some of the data had Nan value. This may pose a challenge later on when I decide to do data cleaning and analysis to seek further insight.

These listings have different kind of Airbnb property. Most of them are Entire home/apartment type (86%) then some private rooms (12%) and only 1% of them are shared room type of property. The average price for an Airbnb listing in Paris is approximately € 111 per night. The dataset has so much of information, and it includes 64 different columns regarding different details. Some of the interesting and useful columns are as follows: if host is super host, host's identity verified or not, no. of bedrooms, no. of bathroom, type of property, availability for 30 days, availability for 365 days, price, no. of guests, neighborhood details, minimum number of nights, if host is responsive, latitude, longitude and if location is exact, user reviews etc.

This information allowed for quick analysis on the data set without much preprocessing. Then we also had a lot of redundant information. For example, the dataset included street names, latitude-longitude on which the housing is located, but it also provided the city, zip code, and general neighborhood name. All of this information can be derived from just the latitude-longitude. The description and the summary provided are also so similar to each other so I can merge some columns of details into onw. Some feature won't add any value to the modelling as they are same for all the listings. As we know the unique the data, the better the data so I chose to remove those columns for processing. Doing this gave a better dataset to work on that will be actually give better results.

This dataset provides answers of the various questions' information about listings such as:

1. What was the price of the listing at a particular day of the year?
2. What is the availability throughout the year?
3. Whether a host is super host or not?
4. What are the number of minimum nights to be booked?
5. How many people a particular listing can accommodate?
6. What is geographical location and if it's verified?
7. What type of room is available?
8. What is the user review?

C. *Exploratory Data Analysis*

To get a good sense of the type of data I am working with and finding a few trends and similarity, I decided to check that by graphs. I began by analyzing how prices correlate with respect to different features such as room configurations (number of bedrooms, room type), day of the year, neighborhood etc. I carried out total exploratory analysis on this dataset. Few graphs and interesting findings are given in brief in following sub sections. There was a total of 64k listings in this dataset.

(a) Price of a room per night: So as the figure shows most of the listings around 100euros and mostly less than that. There are some very expensive and very cheap listings are also there however they are fewer.

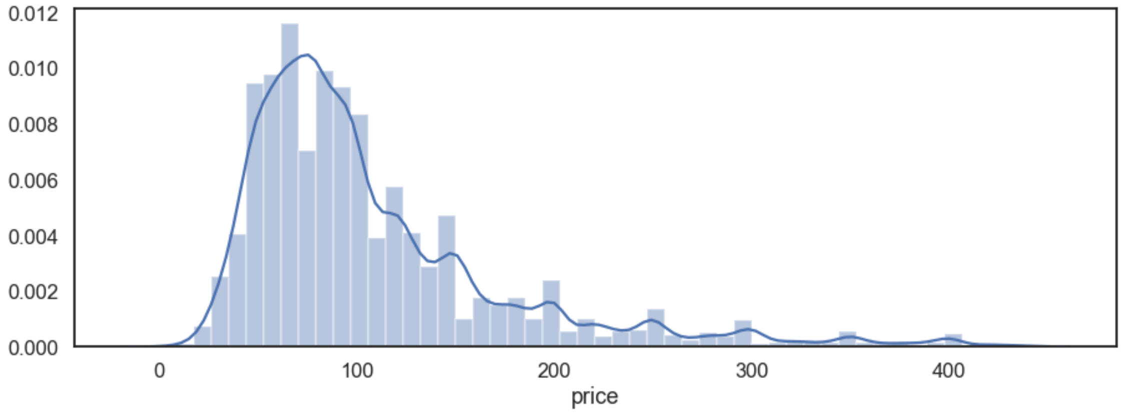

Fig 2 : Price of room per night

(b) Price vs Review ratings: To find out something interesting from the dataset, I did a comparison of price with review rating scores that is as follows:

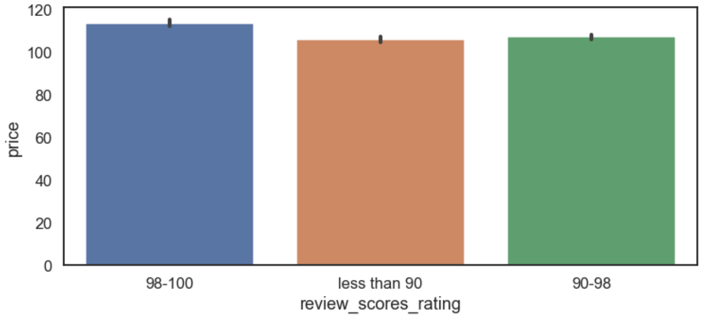

Fig 3 : Price vs Review

Here we can see that there is review ratings are higher for relatively expensive listings. That shows that quality has direct connection with price and hence customer satisfaction. Similarly, another graph can be seen as below that is a heat graph for review and price. I did further analysis as per the findings above.

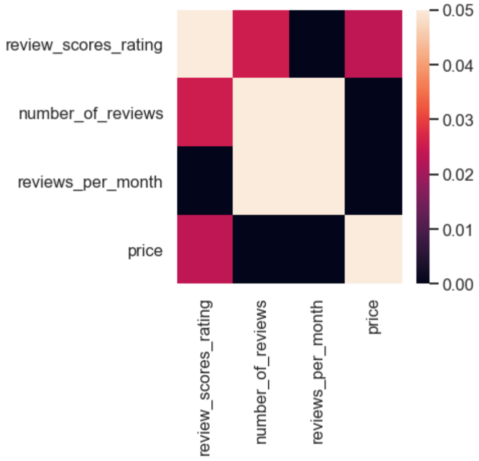

Fig 4 :Price vs Review scores

2. Predictive Task:

There can be two interesting task that can be performed over this data. One of them could be future price setter or price predictor for an Airbnb listing that is totally based on various input features that I think can play important role for price prediction. That includes neighborhood location, type of room etc. Another task is to find out necessary qualities needed in a host that might turn a host into a super host.
So for that I divided the dataset into three sections that is Train data (50%), Test data(25%) and Validation data(25%) before the processing.

Modeling the Price:

The dataset provided a great insight into the features that had a strong influence on how the property would be priced. The model is built on those feature that might affect the price. A few useful features that could be used are, room type, number of bedrooms, number of bathrooms, neighborhood area, no of extra guests and cancellation policy. Excluded the top 1-2 percentile of prices from the prediction tasks as they were big outliers.

Modeling the Super host:

There is a special tag given to some of the hosts at Airbnb as superhosts.These are the hosts, who provide a shining example to other hosts and extraordinary experiences for their guests. A Superhost title provides a number of benefits to customers, Airbnb and the host. So I studies the Airbnb reviews and listings to learn more about important featires to decide for a superhost. Though there are a set of guidelines provided by Airbnb on how to become a super host, we found there are many more latent factors that goes into determining if a host would become a super host. In the study, I visualized various features to determine their relationship with making the user a super host. At the end of the study, I also ran a XGBoost Classifier and Random Forest Classifier to check if the new user is a super host and determined it's qualities for being a superhost.

Important Features to be considered in modeling:

As the dataset has a lot of information, I decided to focus on some of the important features that has some value in changing the end result of this task. So, following are the features considered:

(i) Relevant features for modeling prices:

From the exploratory data analysis of the dataset, it is found that the features related to neighborhood, room type, accommodates, number of bedrooms and number of bathrooms etc. influence the prices the most. Some of the important features are as listed below:
(a) Property type: 26 types of property, ex. Townhouse, Villa, Guesthouse.
(b) Room type: 3 types of rooms mostly, Private Room, Entire home / apartment, Shared room.
(c) Bathrooms: The number of bathrooms that a property has.
(d) Bedrooms: The number of bathrooms that a property has.
(e) Beds: The number of beds in the property.
(f) Bed type: different types of beds, ex. Pull-out Sofa, Couch etc.
(g)Neighborhood: neighborhood of the property.
(h)Is location exact?
(k) Accommodates
(i) Review ratings
(l)Minimum number of nights
(m)host response time etc.

(ii)Modeling super host:

To predict what are the qualities needed for a host to be promoted as a super host, I chose some of those important features.

(a) number of reviews: Number of reviews written by the guests for a host.
(b) Property type: 26 types of property, ex. Townhouse, Villa, Guesthouse.
(c) Room type: 3 types of rooms, Private Room, Entire home / apartment, Shared room.
(d) Bathrooms: The number of bathrooms that a property has.
(e) Bedrooms: The number of bathrooms that a property has.
(f) Beds: The number of beds in the property.
(g) Bed type: different types of beds, ex. Pull-out Sofa, Couch etc.
(h) Reviews scores communication
(i) Host Response Time
(j)Neighborhood: Neighborhood location
(k) Accommodates
(l) Review ratings

3. Model

Here are our experiments with different models for both our predictive tasks.

A. Modeling prices

(a) Baseline: Just a prediction of the mean of all the prices of cleaned dataset
(b) Linear Regression: To analyze the linear relationship between the test and train data.
(b) Ridge Regression: To improve on the previous model so used Ridge Regression which has L2 regularization in it.
(c) Lasso Regression: In cases where relevant information is smeared over large parts of the spectrum asking the regularization to drop variates will low co-efficient is not a particularly sensible approach. Two parameters which are very well correlated maybe dropped by Lasso Regression.
(d) Random Forest: Random forests are an ensemble learning method for regression, that operate by constructing a multiple decision trees at training time and outputting the mean prediction of the individual trees as the final prediction. Random decision forests prevent decision trees' overfitting by optimizing the tuning parameters that governs the number of features that are randomly chosen to grow each tree.
(e) Gradient Boosting: Gradient Boosting is an ensemble technique in which weak predictors are combined in building a better model. These weak predictors learn from the misclassifications from the previous steps and better in the next steps by boosting the importance of incorrectly predicted data points. The aggregate forecast got from each of the weak learners will be much better than each of the learners alone. We got the best performance in the case of Gradient boost regressor.

B. Modeling Superhost

I used two different models to predict if the new user is a super host or not.
Apart from general classification, we used different classifier for modeling superhost that are as follows:

(a) Decision Tree Classifier: Decision tree learning uses a decision tree as a predictive model which maps observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. A decision tree is a simple representation for classifying examples.
In the task of classifying super hosts, there is a single target feature which is whether the host is a super host or not. Each internal node in a decision tree is labeled with an input feature. Each leaf of the tree is labeled with a class.

(a) Baseline models: we implemented two models which we act as baseline for accuracy and precision.
(i) Baseline1: Hosts with reviews/ratings=100 were predicted as superhosts. This gave us our baseline precision
(ii) Baseline2:All hosts were predicted as not a super host . This gave us our baseline accuracy.
(b) Logistic Regression: Similar to linear regression, but instead of predicting the actual value we used logistic functions to predict a probability of 0 or 1.
(c) Random Forest classifier: We used Random forest classifier instead of decision tree as random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
(c) XGBoost Classifier: XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting that solved problems in a fast and accurate way.

4.Literature

The dataset used here is from non-commercial and independent website insideairbnb.com . It provides a set of tools and data that allows to explore how Airbnb is really being used in cities around the world. By analyzing publicly available information about a city's Airbnb's listings, Inside Airbnb provides filters and key metrics so we can see how Airbnb is being used to compete with the residential housing market. Inside Airbnb provides the data for the research purposes open to all.

Previous Work

There is no work found on this particular dataset as this dataset is compiled very recently in September 2019. There were different type on some older dataset but the prediction results were different as well as the goals were different. There is some work done about different US cities as well as the other nation's cities. However, I believe I have added more to it and tried to make my model give better predictions than all the previous work. the work was not There is some prediction work done for like London, Boston, Seattle and older dataset for Paris. I tried to make my work more specific and accurate. The machine learning techniques used in the previous work was linear regression with some limited features like number of bedrooms, bathrooms etc. Hence the error rate was quite higher than I got for this project. Our project is more accurate and reliable because it is trained and tested on quite large dataset. The mean square error is quite better in this project rather than the previous work. The accuracy achieved in this project is a little higher than the other related work done in past. I used and compared all the different important machine learning techniques for the dataset. I performed Exploratory data analysis and tuned the dataset by removing outliers. The previous work was just on the dataset directly provided. There wasn't any further analysis and data cleaning used. This made our results better than the previous work
In the super host prediction, we used two different classifiers and gave a comparative analysis over the result. There were not something similar done for the Paris dataset of Airbnb so I feel this project has a greater value on Airbnb Paris as my findings are quite accurate.

5.Results:

Modeling prices

For this task, I defined some baselines and then worked on the betterment for the better accuracy. As the first step towards building the models, we built the models based on Baseline modeling, Linear Regression, Ridge Regression, Lasso Regression, Random forest and Gradient Boosting. These models were initially built without cross-validation.
Mean Square Error (MSE) was used as a metric to compare the models. As shown in the Table 1, Linear Regression and Ridge Regression gave comparable errors. The Ridge regression was done with different regularization parameter. As the errors of the two models are comparable, the regularization added in Ridge did not impact the model a lot in terms of error rate.
Lasso Regression performed the worst on our model as shown in the Table 1. Reason I can think that model would have dropped many features that could add value to the model as the L1 norm used in Lasso drops the features with lesser significance.
The ensemble techniques like Random Forest and Gradient Boosting performed well and were comparable in performance as shown in Table 1.

After this I did a regression on test and train data as well and shown. As a next step, we checked R^2 statistics for all these regression model. As it is a coefficient of determination, that will give some information about the goodness of fit of a model. The comparison between the models before and after cross-validation (Train and test) are shown in the Table 1.

Model	**Train Error (MSE)**	**Test Error(MSE)**	**R^2 Statistic**
Baseline	4813	5231	0.06
Linear Regression	1721	1728	0.63
Ridge Regression	1627	1656	0.64
Lasso Regression	2045	2056	0.545
Gradient Boosting Regression	1443	1556	0.66
Random Forest Regression	460	1643	0.668

Table 1 : Regression Models with MSE and R^2

GBR and Random forest gave better results, we performed hyper parameter tuning on the validation set to and tested on the test set to confirm our hypothesis.
Tuning Metrics:
For GBR: Number of estimators = 300, learning rate = 0.1 and maximum features = sqrt.
For Random Forest: Number of estimators = 500, depth = 20 and maximum features = sqrt.
This gave some better results as shown in Table 2

Model	**Hyperparameter Tuning**	**Test Error (MSE)**	**R^2 Statistic**

Gradient Boosting Regression	n_estimators=300
	learning_rate=0.1
	max_features='sqrt'
		1443	0.715

Random Forest Regression	n_estimators=500
	depth=20
	max_features='sqrt'
		1354	0.720

Table 2 : Regression Models with MSE and R^2 after hyper-parameter tuning

Apart from that, we checked the features of importance for the price prediction as shown in figure 15. Some important features are:

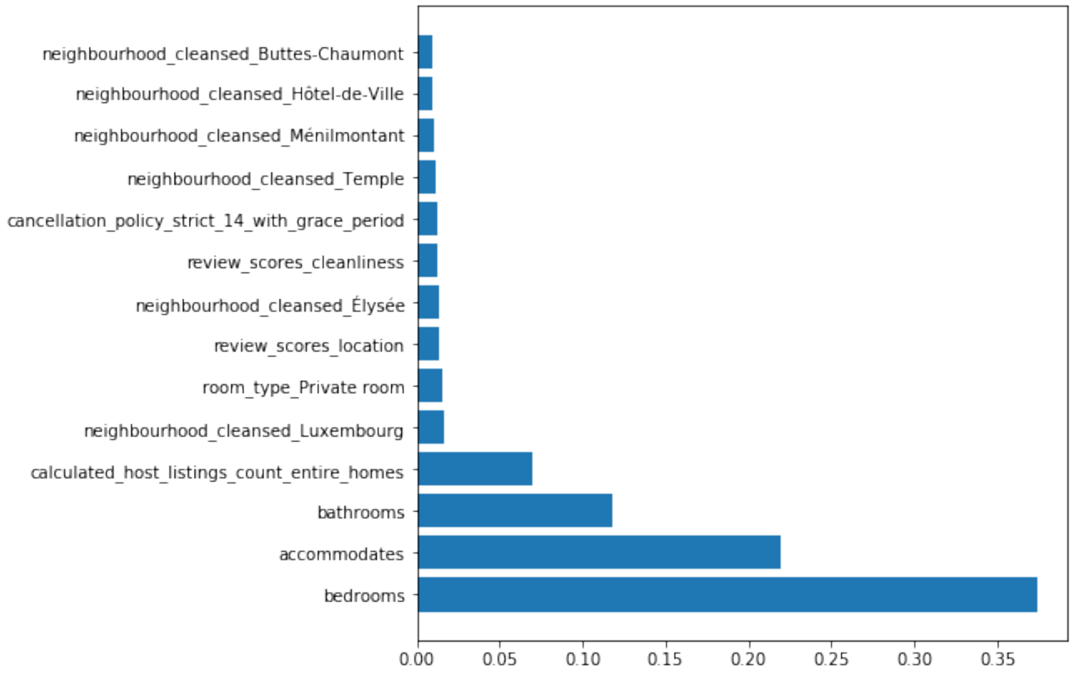

Fig. 5 : Feature of importance for Price prediction task

(a) Room type
(b)Accommodates
(c) bedrooms
(d) bathrooms
(e) Neighborhood etc.

Modeling Superhost:

Model	Train Accuracy (%)	Test Accuracy(%)	Precision
Baseline1(rating =100)	0.49	0.30	0.18
Baseline 2(All zeros)	0.78	0.74	0.0
Logistic Regression	83.30	83.04	0.64
Decision Tree Classifier	97.68	83.54	0.44
Random Forest Classifier	97.54	85.34	0.7
XGBoost Classifier	85.63	84.3	0.68

Table 3: Classifiers with Accuracy and Precision Results

The problem of modeling a super host is rather challenging. The major challenge being that there are not many distinguishing parameters between the super hosts and the normal hosts. To deal with this challenge, we performed exploratory analysis and found the various features that were distributed differently between the two type of hosts.

To deal with the problem of imbalanced dataset we experimented with using Baseline with rating 100. The accuracy was very low in that case so checked for another baseline that is choosing all zeros. This was done just to get us baseline accuracy as its precision would obviously be zero. To improve on our accuracy and precision, we experimented with logistic regression and decision trees. We got a good improvement in the accuracy as well as precision. We got 97.68% training accuracy and 83.54% testing accuracy for decision tree that might be due to overfitting issue. So tried random forest that gave a slight improvement on accuracy\ and precision and that was the best result. Then I tried using XGBoost, the accuracy was better in the case of test accuracy.

XGBoost and Random forest gave better results, we performed hyper parameter tuning on the validation set to and tested on the test set to confirm our hypothesis.

Tuning Metrics:

For XGBoost: Number of estimators = 100, gamma = 0.01 , max depth= 5, min child weight=6 and min leaves=2
For Random Forest: Number of estimators = 100, max depth = 20 ,maximum features = auto and criterion=gini

This gave me better results as shown in Table 4

Model	**Hyperparameter Tuning**	**Test Accuracy (%)**	**Precision**

Random Forest Classifier	n_estimators=100
	max_features=’auto’
	max_depth=8
	criterion=’gini’
		89.54	0.76

XGBoost Classifier	n_estimators=100
	max_depth=5
	min_child_weight=5
	gamma=0.01
	min_leaves=2
		90.48	0.77

Table 4: Best performing Classifiers with Accuracy and Precision Results

To find out the qualities needed for a host to be a superhost, there are a number of features that come into play. Airbnb lists a couple of attributes that are paramount for becoming a super host. I listed out some impotant qualities as follows:

_ Number of reviews written by the guest for a host
_ Level of cleanliness of the host
_ Responsiveness of the host
_ How accurate is the listing
_ Review score rating
_ Type of the room provided

The more are shown in the figure according to their importance value:

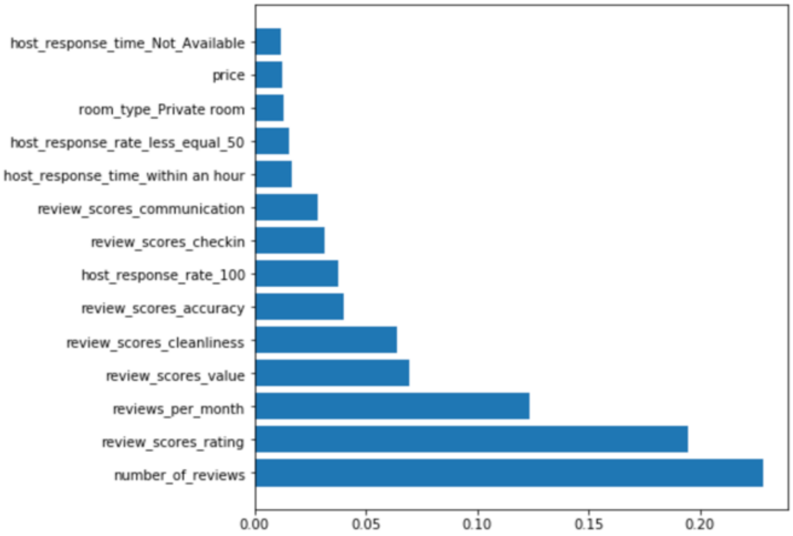

Fig. 6 : Feature of importance for Superhost predicrion task

This project had two different tasks and both of them gave good results .First, the features that influenced price of the listings and the Second, qualities that make a super host.
For modeling prices, we studied various models. Machine learning techniques like Random Forest and Gradient Boosting proved to be better than most other models.
To derive characteristics associated with a super host, I performed exploratory data analysis to find that cleanliness, location, communication, number of reviews were major characteristics through which I can distinguish a super host. Then all this got verified by building classifiers such as XGBoost and trees, and found our results were appropriate.