Locating the best Airbnb to stay at when visiting Singapore

Pua Wan Xin

This report is for the IBM Applied Data Science Capstone Project

Introduction:

1.1 Background

Singapore is a small city country that has developed a reputation for being a 'Garden City'. It has a built environment that has many charms due to its rich multi-racial culture. For example, Singapore consists of 3 major ethnic groups, Chinese, Malays and Indians, and has cultural enclaves Chinatown, Little India, and Arab Street, where one can have a taste of diverse cultures and festivals. Tourism in Singapore represents a big part of its economy, with a total of 18.5 million international visitors spending a total of SGD27.1 Billion in 2018 (Tiffany Tay, 2019). In order to attract visitors, Singapore has built many attractions, the latest one being Jewel mall at Changi Airport, completed in Oct 2019. Jewel encapsulates Singapore as a garden city as it is nature-themed, with the world's tallest indoor waterfall, surrounded by a terraced forest with plants from around the world.

1.2 Problem

As Singapore takes importance on tourism, the number of hotels and places to stay at whilst visiting Singapore is numerous. This poses a question for tourists – Where to stay?

1.3 Aims and Objectives

Airbnb is one of the popular platforms that people use in order to find accommodations. In order to answer this question of where to stay, this project aims to find the best place for visitors to stay in.

To find the best location to stay at:

1) Divide Singapore into regions
2) Agglomerate attractions and venues for tourists
3) Find the best regions to stay at with the most attractions
4) Find the best Airbnb listings in the identified regions

This project will also study Airbnb prices to see what are the determinant factors – how neighborhood plays a part, and construct a model that can predict prices. Abhinav Sagar (2019) provides an explanatory guide on how to do this.

Data Sets

2.1 Singapore Neighborhoods

2019 Planning Area Boundary [Urban Redevelopment Authority]: Data downloaded from data.gov.sg (https://data.gov.sg/dataset/master-plan-2019-planning-area-boundary-no-sea).

Data type: GEOJSON

This Dataset will be used to find the Neighborhood (Neighborhood) that has the most attractions for visitors.

2.2 Attractions in Singapore

The attractions datasets will be used to find attractions in Singapore. The locations of the venues will be agglomerated to the Neighborhoods of Singapore

1) Tourist Attractions [Singapore Tourism Board, STB]: Data downloaded from data.gov.sg (https://data.gov.sg/dataset/tourist-attractions).

Data Type: KML

2)Foursquare API

Get Venue Recommendations API. Documentation: https://developer.foursquare.com/docs/api/venues/explore

The Arts and Entertainment category will be used as it is most applicable for tourists

2.3 Accommodation - Airbnb

Airbnb Singapore Listing dataset: Data downloaded from http://insideairbnb.com/get-the-data.html

Airbnb is the type of accommodation chosen in this project for tourists to stay at. Airbnb listings data, last extracted 5 December 2019 will be analyzed

Locating best place to stay

Methodology

3.1 Data Processing and Cleaning

To process the Neighborhoods of Singapore, the geojson was loaded in python using pandas and converted into a table readable format. The center points of each Neighborhood were also found by using the geopandas library. The table fields for the Neighborhood are:

Neighborhood	Geometry	Latitude (centerpoint)	Longitude (centerpoint)

The center points of each Neighborhood were then used to find foursquare venue recommendations. An iteration was performed with each lat, lon center points of the Neighborhood to find arts and entertainment related venues within a 1km distance from the center points. The table fields for the venues are:

Neighborhood	Neighborhood latitude	Neighborhood Longitude	Venue name	Venue Latitude	Venue Longitude	Venue Category

For the Data.gov.sg attractions dataset, the kml file was converted into a geojson and parsed using geopandas to a table readable format. The name of the place had to be cleaned. The original format of the place name was within a link, for example, "www.yoursingapore.com/en/see-do-singapore/culture-heritage/heritage-discovery/chinatown-heritage-centre.html". The link was split in order to extract only the name (e.g Chinatown Heritage Centre). This was done for each row. A spatial join was then performed with the attraction dataset with the neighborhood dataset to find the neighborhoods of each of the attraction locations. The table fields for the attractions are:

Neighborhood	Neighborhood latitude	Neighborhood longitude	Venue name	Venue latitude	Venue longitude	Venue geometry

To process and clean the Airbnb data, firstly, the average price per person per night of each listing was calculated (data fields: 'Price'/'Accommodates'). The next step was to filter the listings for those suitable for tourists. To do this, the minimum nights was set to 1, and maximum nights to more than 6 nights.

3.2 Data Analysis

3.2.1 Finding the neighborhoods with the most attractions

After generating the list of venues with their respective neighborhoods, the top 10 neighborhoods with most venues from the STB (Singapore Tourism Board) dataset were found, and the neighborhoods with 50 venues in the list from the foursquare API were found. Those with 50 venues were used as 50 is the limit that the foursquare API can return, therefore those that hit the limit were used.

STB venues	
Neighborhood	Count
Downtown Core	29
Bukit Merah	11
Rochor	11
Outram	10
Museum	7
Southern Islands	6
Central Water Catchment	4
Queenstown	4
Marine Parade	3
Marina South	3

Foursquare API venues	
Neighborhood	Count
Museum	50
Marina South	50
Downtown Core	50
Bukit Batok	50
Outram	50
Orchard	50
River Valley	50
Newton	50
Singapore River	50

The common neighborhoods identified by both datasets were Downtown Core, Marina South, Museum, and Outram. To carry out further analysis, the neighborhoods that were identified in either of the datasets were used.

Table 1: Identified neighborhoods best to stay in when visiting Singapore

Museum	Marina South	Downtown Core	Bukit Batok	Outram
Orchard	River Valley	Singapore River	Bukit Merah	Newton
Queenstown	Central Water Catchment	Southern Islands	Marine Parade	Rochor

The total number of listings that fit the criteria for suitability of tourist stay, and within the identified neighborhoods (table 1) is 952.

3.2.2 Finding out the cost of staying in the identified neighborhoods

Using the list of neighborhoods identified (table 1), analysis of the price of the neighborhoods was carried out.

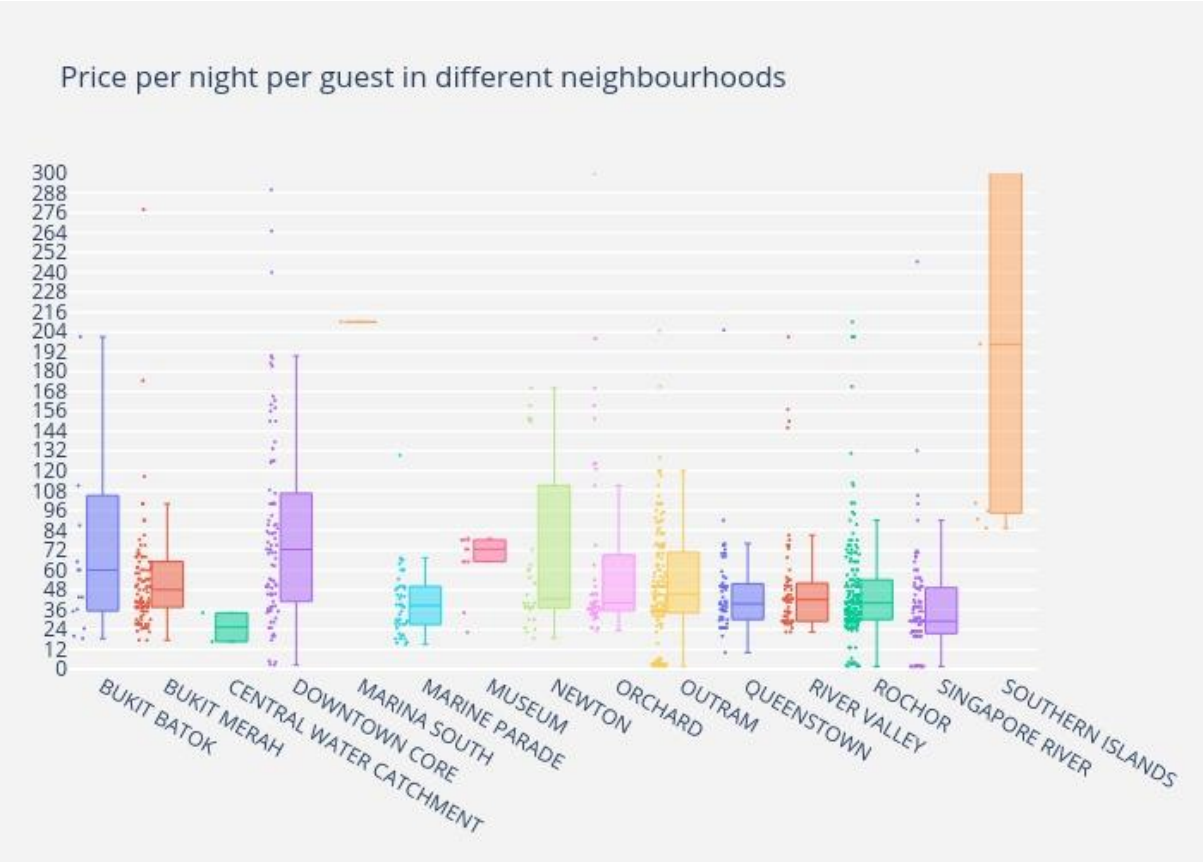


Fig 1: boxplot showing price per night per guest of identified neighborhoods. To view interactive version: https://plot.ly/~employmee/1/

3.2.2 Analyzing popularity against price

A plot of average number of reviews against the price of the identified neighborhoods was created to identify neighborhoods that tend to have the neighborhoods that are both affordable and popular.

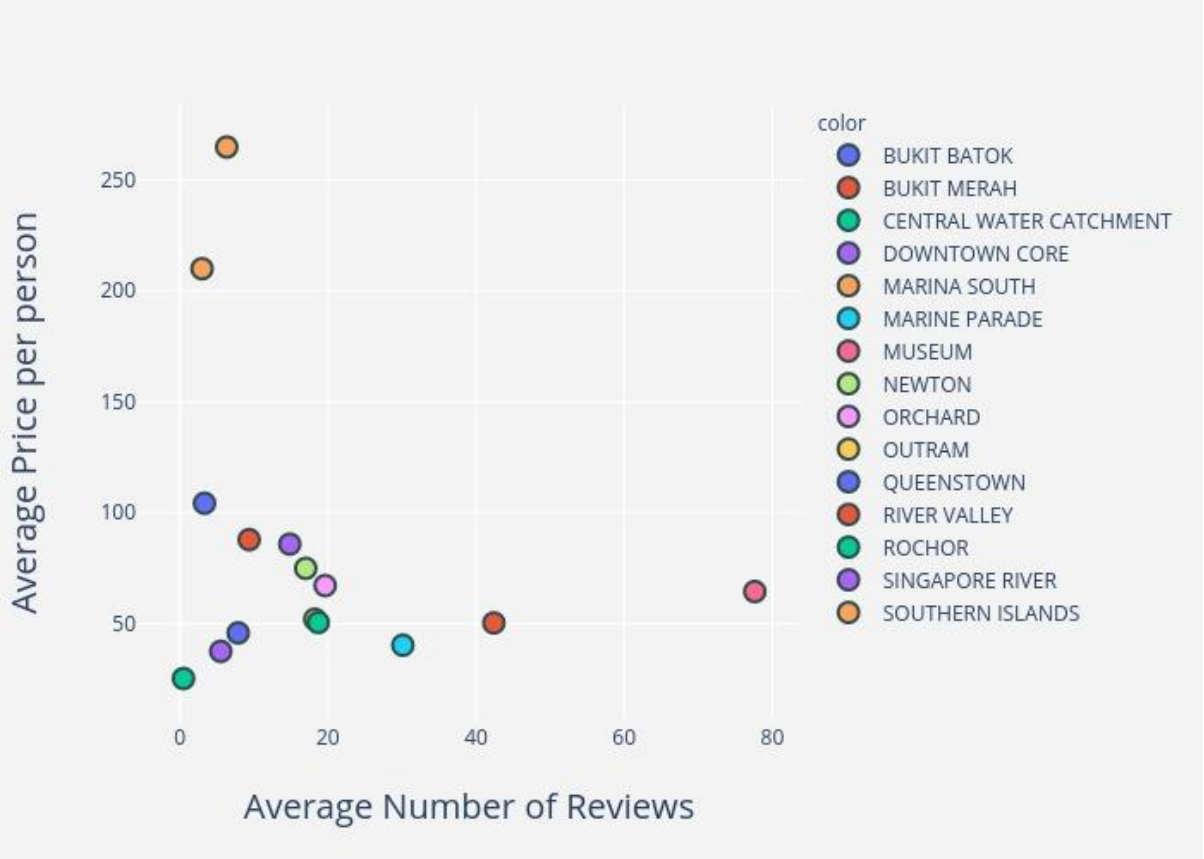


Fig 2: scatterplot of price against number of reviews. To view interactive visualization: https://plot.ly/~employmee/8/

The neighborhoods that are popular are River Valley, Museum, and Marine Parade. These 3 neighborhoods are also affordable as they fall under 80SGD per night.

3.2.3 Find the best Airbnb listings

In order to locate the best listings, the listings were filtered on 3 conditions: number of reviews (descending order), review score rating (descending order) and price (ascending order). Fig 3 shows the top 20 listings following the specified queries, and the attractions from data.gov.sg.

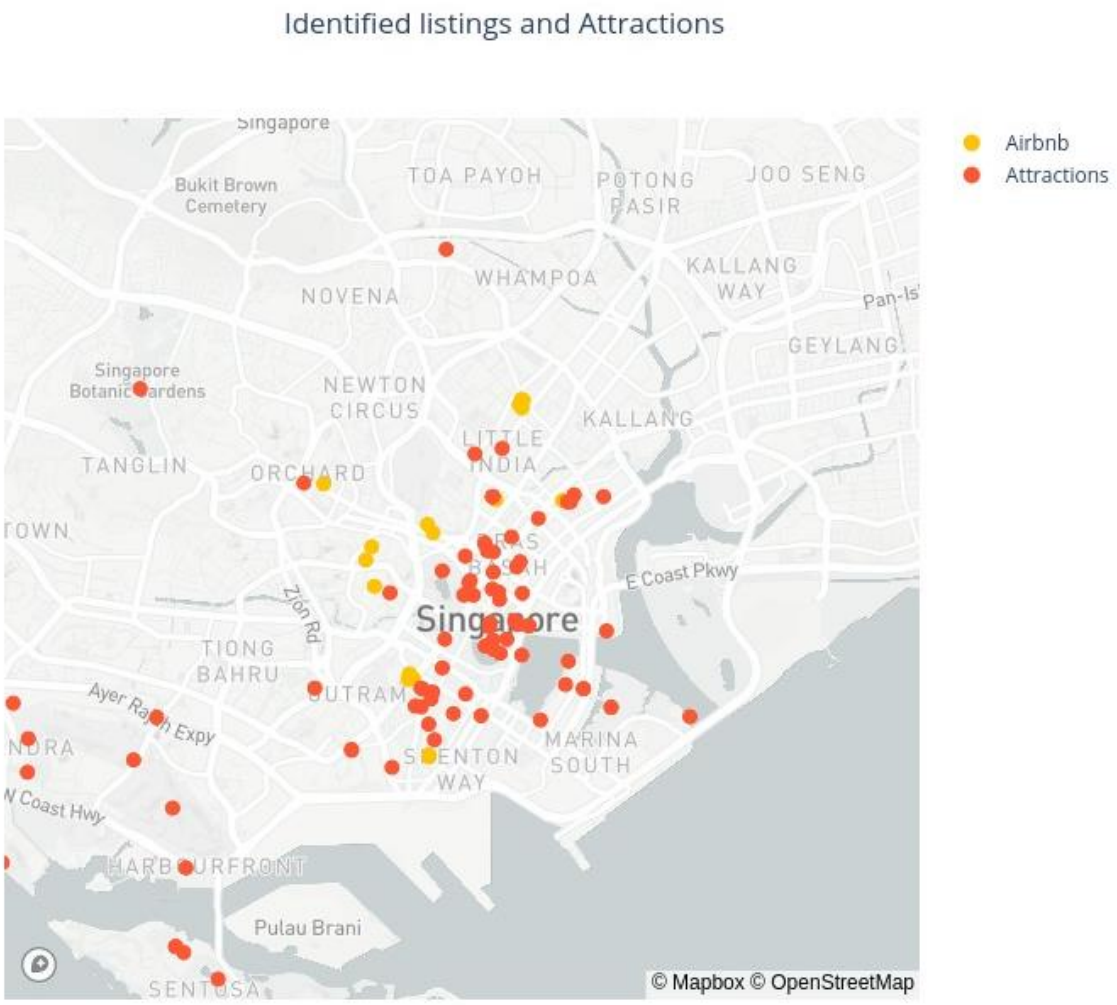


Fig 3: Scatter map showing identified best listings and Singapore attractions. Interactive visualization: https://plot.ly/~employmee/10/

As seen in fig 3, since identified listings were first filtered according to the best neighborhood, they are all located near to where the bulk of attractions in Singapore are found. The interactive visualization link will allow users to hover and see the details of the listing, including the price, number of reviews, review score rating and ID of the Airbnb, and also the name of the attraction.

Predicting Airbnb Price

Methodology

4.1 Selecting Features

First, a set of features from the Airbnb Listings dataset was selected. The features selected are shown in table 2.

Table 2: Selected features from Airbnb listing dataset to check against price

Numerical Columns				
Categorial Columns				
Cancellation policy	Is location exact	accommodates	Reviews per month	Property type
Host total listings count	Extra people	Calculated host listings count	Guests included	price
Review scores rating	Host response time	Number of reviews	Host is superhost	Room type
Instant bookable	Host identity verified	Host has profile picture	Security deposit	Host listings count
Neighbourhood cleansed	Minimum nights	bathrooms	Require guest phone verification	Cleaning fee

In order to clean the missing data from the dataset, missing categorial data were filled with the most often (mode) value, and missing numerical data were replaced with the median value.

4.2 Correlation

A correlation matrix was then created to see how features affect one another (fig 4).

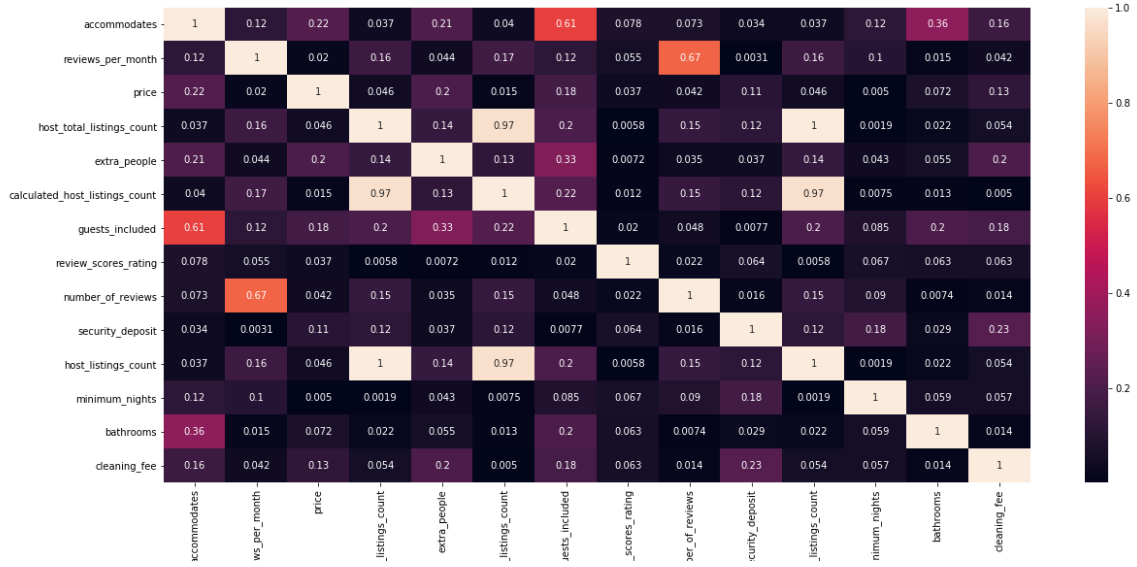


Fig 4: Correlation matrix

Table 3: Correlation to price. From highest to lowest

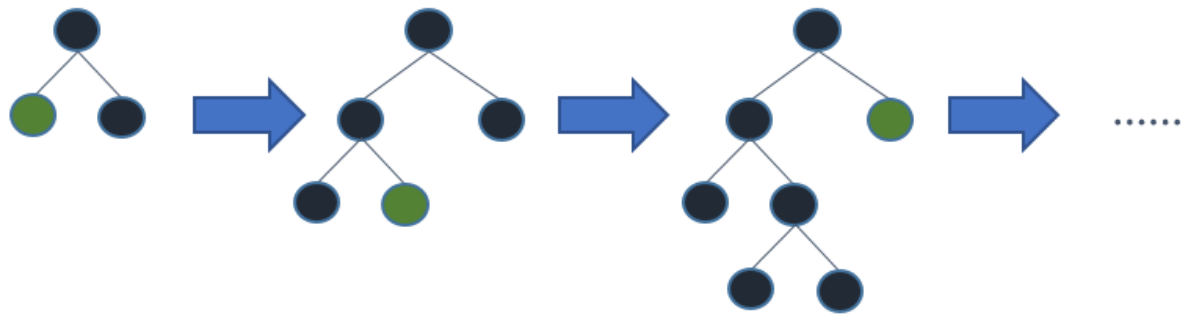
Feature	Correlation
Accommodates	0.216972
Extra people	0.197937
Guests included	0.180531
Cleaning fee	0.129200
Security deposit	0.110909
Bathrooms	0.072492
Host listings count	0.045709
Host total listings count	0.045709
Number of reviews	0.041662
Review scores rating	0.037190
Reviews per month	0.019514
Calculates host listings count	0.014922
Minimum nights	0.00500

The feature that has the highest correlation to price is the number of people that the Airbnb accommodates. In general, the features do not exhibit a high correlation to each other. No significant correlation was also found for any of the features to price.

4.3 Modelling

4.3.1 Methodology

In order to model, the categorical columns will first be transformed, with unique values transformed into columns. Light GBM model was used to train and model price. GBM (Gradient Boosting Deision Tree) is a machine learning algorithm that is widely used due to its efficiency, accuracy, and interpretability (Ke et al., 2017). LightGBM has several advantages such as, higher efficiency and faster training speed, usage of lower memory, better accuracy, support parallel and GPU learning, and can handle large scale data and features. LightGBM grows tree vertically; meaning that it grows tree leaf-wise compared to level-wise. The growth is dependent on choosing the leaf that has a max delta loss. Leaf—wise growth can reduce more loss compare to level-wise (Kirti Bakshi, 2018).



Leaf-wise tree growth

The model was set to train 20% of the dataset, and random state was set to 32 for optimized model.

4.3.2 Results

A plot was made to see how the model was able to predict the price vs the actual price (fig 5).

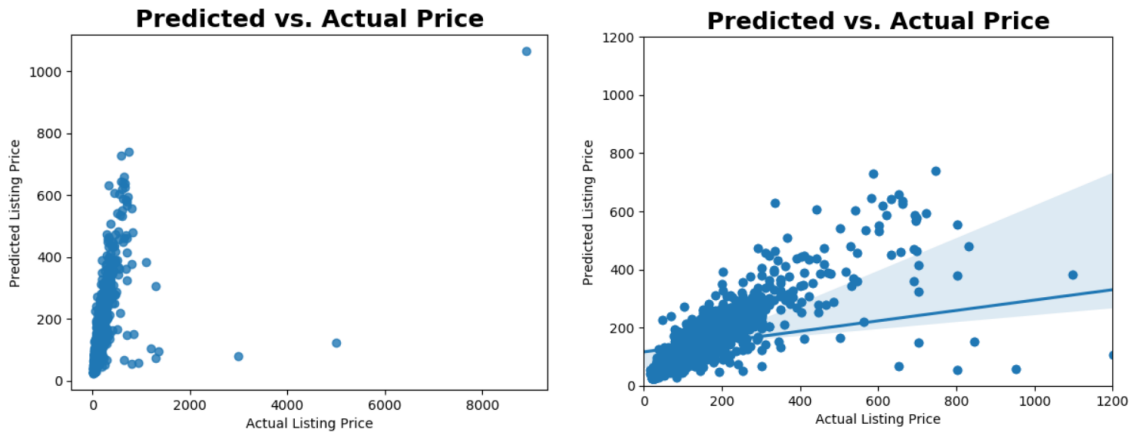


Fig 5: Predicted vs Actual Price (left) without limits (right) set x and y limit to 1200

The model was produced had an R squared score of 0.755.

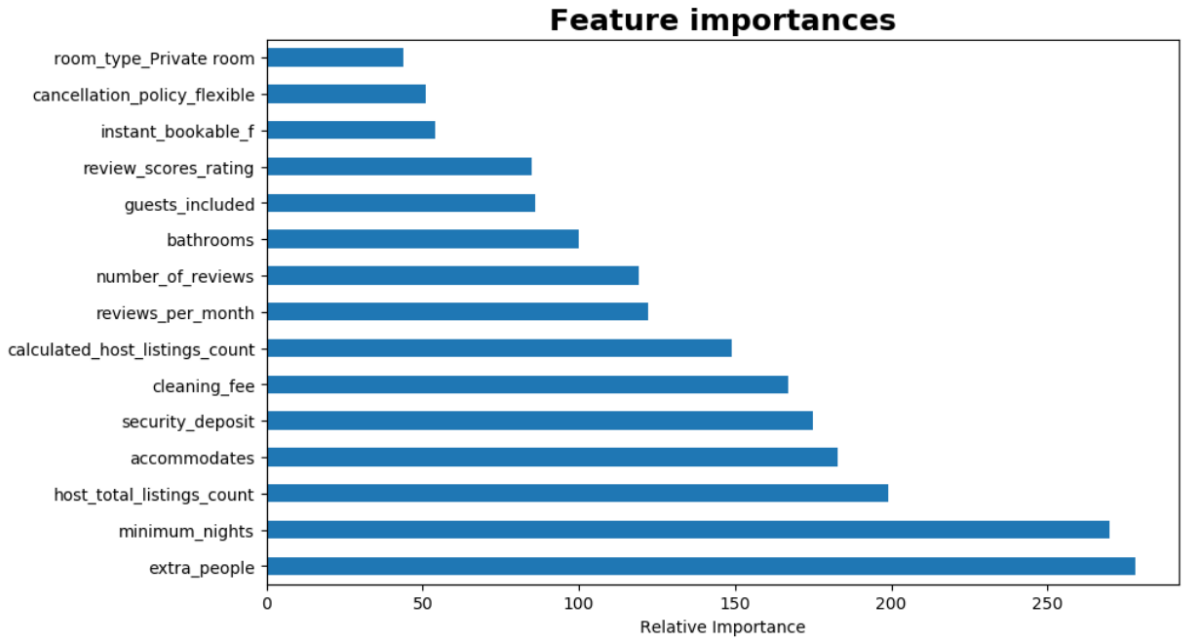


Fig 6: Importance of features to determining price

The results show that extra people are of highest importance. However, inspection of the dataset show that most people tend to put the same number in extra people and accommodates. This means that some people treat the columns to be the same whilst others may treat them differently, causing inaccuracy.

Conclusion

The study found that the best neighborhood to stay in Singapore when visiting would be Museum, River Valley, and Marine Parade, based on Airbnb prices and amount of activities to do for tourists. This study also found that when it comes to Airbnb prices, there is not a certain feature that has a strong correlation to it.

Future Directions

If hotel datasets were readily available for Singapore, it would be much better to use those as hotels are probably more commonly used by tourists when finding a place to stay. Better regressions could probably be found with price as hotels have a way to determine the pricing of their rooms such as number of stars, facilities, location etc. The amenities in Airbnb listings (e.g wifi) were not accounted for in the predicting study and this could be one of the more important features in relation to price. This study could also be improved upon if there were tourist information (E.g how many people came together, age, sex, amount they spent, number of days in Singapore) and where they chose to stay at. A categorial modelling could also then be done to assign the type of tourists to where they would tend to stay.

Reference:

ABHINAV SAGAR *Predicting Airbnb Prices Using Machine Learning in Vancouver* [Online] Medium Published 15 Nov 2019 https://towardsdatascience.com/predicting-airbnb-prices-using-machine-learning-in-vancouver-1b42ca52eece

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146-3154).

KIRTI BAKSHI *LightGBM: A light Gradient Boosting Machine* [Online] TechLeer Published 25 Feb 2018 https://www.techleer.com/articles/489-lightgbm-a-light-gradient-boosting-machine/

TIFFANY FUMIKO TAY *Tourist arrivals, spending in Singapore at record highs* [Online] Straits Times Published 14 Feb 2019 https://www.straitstimes.com/singapore/tourist-arrivals-spending-at-record-highs