Group 18 - Jisoo KIM, Daoqing SU, Ho Ming Anthony WONG, Jonathan Zi-him YORK, Shuqing ZHAN, Tianliang ZHANG

Project Marvel – Turning hosts to SUPERhosts

12 December 2022

Executive Summary

Airbnb, Inc ("**Airbnb**") is a well renowned short-term rental platform and has disrupted the lodging sector across the world with its innovative technology. By leveraging the global sharing economy, Aribnb, Inc has grown to have over 4 million hosts on their platform and have welcomed over 1 billion guests. As Airbnb's business model is primarily on taking commission on both hosts and guests, it is in their best interest to create the best possible hosts and experience for the guests. In order to ensure that the Airbnb guests get the best possible experience, Airbnb concocted the key success factors which can turn any host, into a "Superhost". Hong Kong has a relatively lower number of "Superhosts" and as such, there is an opportunity to further penetrate this market. Our key objective is to find out how we can help Airbnb create more "Superhosts" in Hong Kong.

Logistic regression model has been constructed to identify contributing factors for hosts to become "Superhosts". We found that the number of available amenities and the hosts' acceptance rate as well as the existence of neighbourhood overview have a positive impact on becoming "Superhosts". On the other hand, a slower response time by hosts will negatively impact the chance of hosts becoming "Superhosts". However, whether a listing is bookable instantly appears to have a negative impact for listings in Hong Kong, suggesting that this may not be an important feature to guests looking for a place in Hong Kong.

Additionally, we run Random Forest Regression and XGboost to verify the importance of each variable from another perspective from the logistic regression model and identify any additional contributing variables. It is concluded that hosts having verification emails and slow response time are also important variables explaining the likelihood of becoming a "Superhost".

A further logistic regression model in respect of amenities has been built to analyse what amenities contribute positively to becoming "Superhosts". We found that the availability of the following amenities has a positive impact towards becoming "Superhosts": (i) shampoo, (ii) iron, (iii) hot water kettle, (iv) first-aid kit, (v) coffee maker, (vi) television, (vii) cable television, (viii) dedicated workspace, (ix) dryer, (x) dishes and silverware, (xi) fire extinguisher, (xii) basic cooking utensils, (xiii) hair dryer, (xiv) hangers, (xv) allowing long term stays, (xvi) carbon monoxide alarm and (xvii) locks for bedrooms' doors.

Specifically, (i) shampoo, (ii) iron and (iii) television are the top-three amenities that have the strongest impacts towards becoming "Superhosts".

Both Python and R were utilised to cross-validate the results and compare the pros & cons of using different two programming languages.

Introduction

Airbnb, Inc. engages in the management and operation of an online marketplace between travellers and homeowners. Its marketplace model connects hosts and guests online or through mobile devices to book spaces.[i] Its success highly depends on superb hosts who open their homes to guests from around the world. In 2009, Airbnb launched its first version of the "Superhost" programme.[ii] Throughout the years, Airbnb continues to update its "Superhost" programme. Airbnb sees "Superhosts" as those who go above and beyond in their hosting duties and is a shining example of how a "Host" should be.[iii] "Superhosts" are identified by the badge that appears on their listing and profile. [iv]

As of today, to qualify as a "Superhost", a listing owner must own an account in good standing who has met the following criteria in the past 12 months:

1) Completed at least 10 trips or 3 reservations that total at least 100 nights;
2) Maintained a 90% response rate or higher;
3) Maintained a less than 1% cancellation rate, with exceptions made for those that fall under Airbnb's extenuating circumstances policy; and
4) Maintained a 4.8 overall rating.[v]

The evaluation is conducted by Airbnb on a quarterly basis. [vi]

Being a "Superhost" has immense benefits, including having more visibility from prospective guests, additional earning potentials, exclusive rewards and getting priority support from Airbnb. From a financial standpoint, "Superhosts" notice a 5% increase in weekly views of their listing. They also enjoy an 81% higher occupancy rate compared to normal hosts and earn 60% more daily revenue than regular hosts on average.[vii] Given majority of Airbnb's revenue comes from charging guests and hosts a service fee, Airbnb is also incentivised to encourage more hosts to become "Superhosts".[viii]

Globally, about 19.4% of hosts are "Superhosts". [ix] However, based on the data made available by Airbnb[x], only 10.5% of hosts in Hong Kong are "Superhosts".[xi] As Hong Kong gradually lifts its travel restrictions imposed due to COVID-19, it is prime time to conduct analysis on the performance of Airbnb's listings in Hong Kong and provide insights to hosts in Hong Kong to better prepare the expected influx of travellers, so that they could recoup the grounds lost in the past few years.

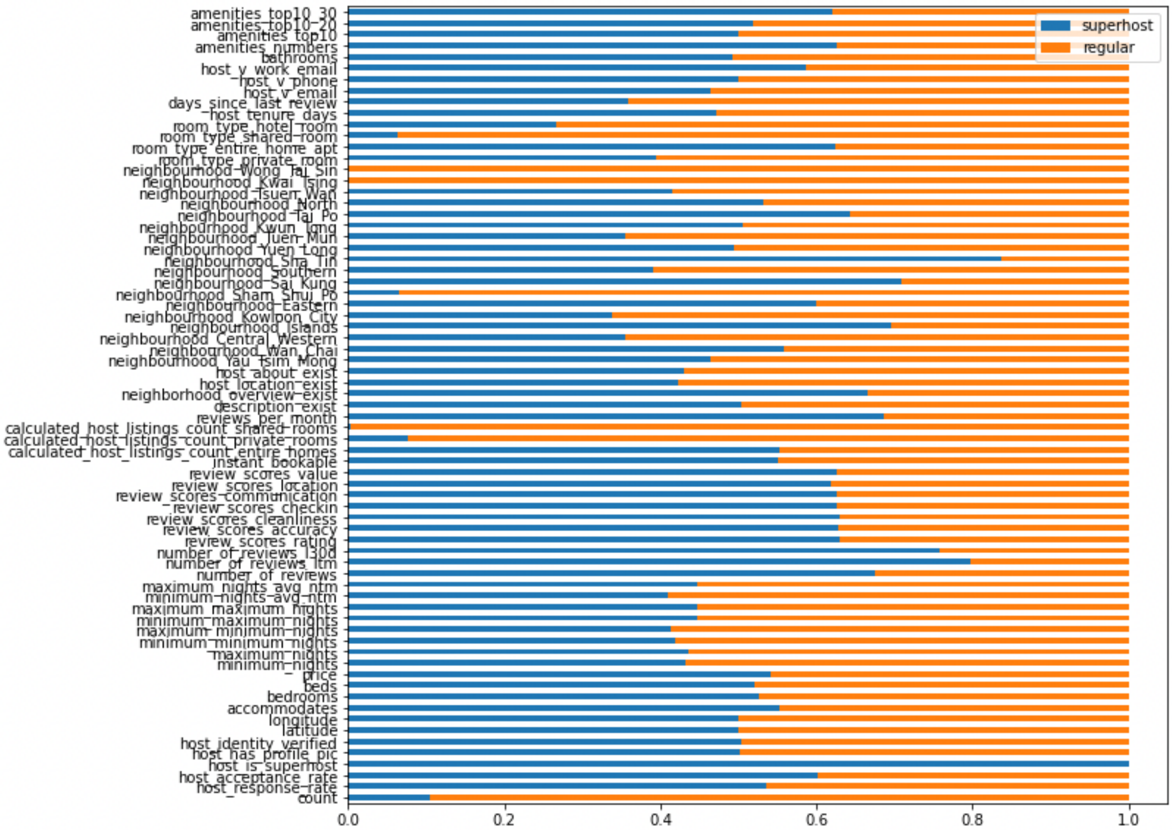
This report provides quantitative analyses on the influential factors that help hosts in Hong Kong to become "Superhosts". In particular, we have conducted logistic regression to analyse the contributing factors. We also specifically conducted logistic regression on the amenities that are typically available in listings in Hong Kong to identify the more important amenities that hosts should consider equipping their listings with in order to become "Superhosts".

Coupled with deep market knowledge, this report provides practical tips to Airbnb in helping aspiring hosts and existing hosts become "Superhosts".

Analyses

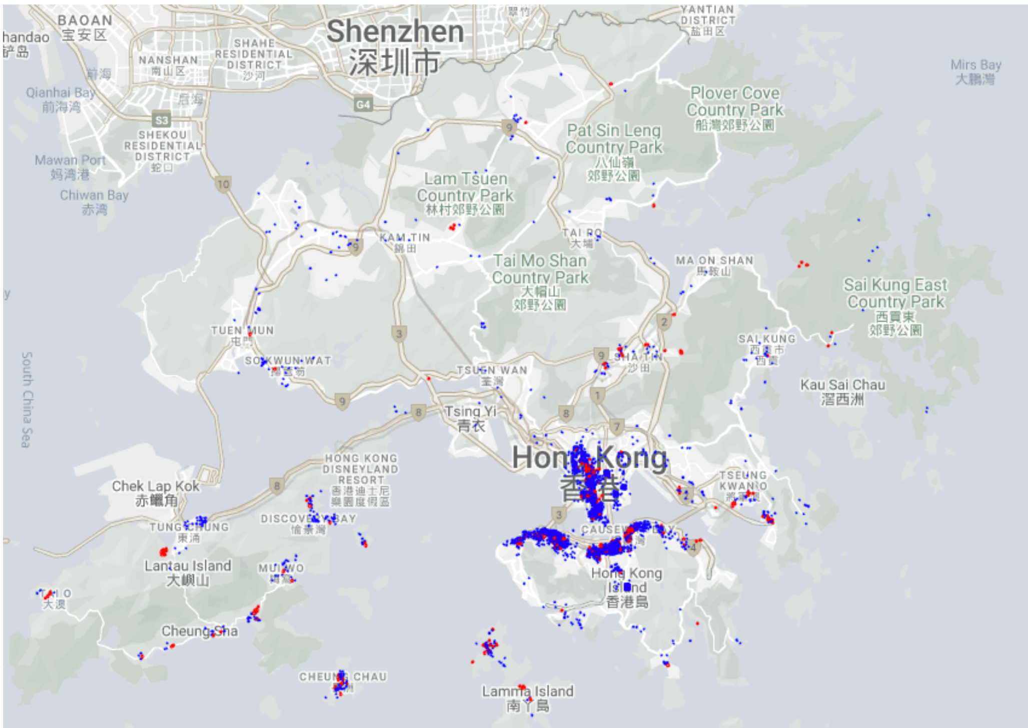
Initial observations

As a start, we have conducted a comparison between regular hosts and "Superhosts" to understand their performances on different attributes. For visualisation purposes, the average values of regular hosts and "Superhost" in respect of each attribute were converted to a relative ratio and visualised in the form of bar histograms. The results are as follows:



As the above chart demonstrates, there are a number of notable differences between the performances of regular hosts and "Superhosts", suggesting that these may be factors determining whether a host can become "Superhost". For example, the more reviews the hosts received, the better the chance of such hosts becoming "Superhosts". Number of total amenities, number of top 10 amenities as well as listing the entire home also appear to be some of the contributing factors. In our models, we seek to prove that number of amenities is indeed a very significant factor in determining whether a host can become a "Superhost".

Meanwhile, from a location perspective, listings in Shatin, Sai Kung and the Hong Kong Island more often than not propel hosts to become "Superhosts", whereas listings in Kwai Tsing, Wong Tai Sin and Sham Shui Po are made by hosts with the lowest rate of "Superhosts". This suggests that better neighbourhood environments may result in a better chance of achieving the "Superhost" status.



That said, achieving a "Superhosts" status in Hong Kong does not automatically translate to a significant increase in revenue. The above result shows that the average price for listings made by "Superhosts" is only marginally higher than those made by regular hosts.

Based on the above initial observations, we have constructed the models below to examine closely the contributing factors to becoming a "Superhost" from a quantitative perspective.

<u>Data Cleaning and Preparation</u>

Before building any models, it is vital to clean and better prepare the dataset made available to us. This process includes: (1) converting data with 'object' data type to appropriate data types; (2) dropping columns that have more than 75% missing values; (3) filling missing data with appropriate entries; (4) dropping variables; and (5) creating categorical/dummy variables.

In relation to (4) above, we dropped variables that (i) cannot be easily actioned upon by hosts or (ii) are not relevant to the business questions we aim to answer in this project. For example, we dropped all url related columns such as 'host_picture_url', 'host_name', 'latitude', and 'last_scraped' as these variables contain no relevant information in our analysis. We also dropped variables which are not pragmatic for hosts to act upon such as the location of the property, the location of the hosts, the number of toilets and bedrooms the property has and the type of the property, however these variables may be relevant.

In relation to (5) above, we have created a number of categorical variables for 'host_response_time'. We have also created a number of variables in respect of the amenities available in respect of each listing, which were initially provided as texts in the initial data sets. For example, we created a variable of 'amenities_numbers' for counting the number of amenities for each listing, and broke down the amenities into dummy variables.

Models

Model 1: Overall Logistic Regression

The first model we constructed is logistic regression, with 'host_is_superhost' being the dependent variable, with the aim to analyse the features that are most important in helping hosts to become "Superhosts" in Hong Kong.

To make our recommendations as concise as possible, we undertook a model reduction exercise using the AIC Forward Selection method to further narrow down the number of variables to be used in the logistic regression. The following results identify the more important variables to be used in our final logistic regression model.

Generalized Linear Model Regression Results

Dep. Variable:	host_is_superhost	**No. Observations:**	4050
Model:	GLM	**Df Residuals:**	4038
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-931.24
Date:	Sun, 11 Dec 2022	**Deviance:**	1862.5
Time:	18:30:38	**Pearson chi2:**	2.81e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-4.3479	1.132	-3.839	0.000	-6.567	-2.128
amenities_numbers	0.0701	0.006	10.928	0.000	0.058	0.083
host_acceptance_rate	2.7821	0.243	11.469	0.000	2.307	3.258
host_v_email	-1.9203	0.196	-9.787	0.000	-2.305	-1.536
neighborhood_overview_exist	1.1231	0.144	7.787	0.000	0.840	1.406
host_identity_verified	-0.8735	0.142	-6.148	0.000	-1.152	-0.595
response_time_a_day	-1.5499	0.301	-5.157	0.000	-2.139	-0.961
response_time_a_few_days	-1.9440	0.597	-3.254	0.001	-3.115	-0.773
beds	-0.1910	0.050	-3.786	0.000	-0.290	-0.092
instant_bookable	-0.5914	0.147	-4.016	0.000	-0.880	-0.303
host_about_exist	-0.2985	0.137	-2.184	0.029	-0.566	-0.031
host_has_profile_pic	1.6141	1.109	1.456	0.145	-0.559	3.787

The exercise above is followed by conducting a variance inflation factor (VIF) analysis to identify any existence of multicollinearity issues, where we dropped the variable 'host_has_profile_pic' at this stage, which seemed to cause significant multicollinearity issues.

	variables	**VIF**
0	amenities_numbers	5.237735
1	host_acceptance_rate	3.853621
2	host_v_email	17.219852
3	neighborhood_overview_exist	2.294009
4	host_identity_verified	2.728111
5	response_time_a_day	1.161903
6	response_time_a_few_days	1.102534
7	beds	2.695326
8	instant_bookable	1.845464
9	host_about_exist	5.039361
10	host_has_profile_pic	20.956781

	variables	**VIF**
0	amenities_numbers	5.072229
1	host_acceptance_rate	3.780598
2	host_v_email	7.852591
3	neighborhood_overview_exist	2.290653
4	host_identity_verified	2.721392
5	response_time_a_day	1.157507
6	response_time_a_few_days	1.091049
7	beds	2.661999
8	instant_bookable	1.844996
9	host_about_exist	4.616332

At the end, our final logistic regression model contains 9 variables. And the final output is as follows:

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:       host_is_superhost   No. Observations:                 4050
Model:                             GLM   Df Residuals:                     4039
Model Family:                 Binomial   Df Model:                           10
Link Function:                   logit   Scale:                          1.0000
Method:                           IRLS   Log-Likelihood:                -932.70
Date:                 Sun, 11 Dec 2022   Deviance:                       1865.4
Time:                         18:30:38   Pearson chi2:                 2.81e+03
No. Iterations:                      7
Covariance Type:             nonrobust
===============================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                     -2.7300      0.216    -12.648      0.000      -3.153      -2.307
amenities_numbers              0.0698      0.006     10.891      0.000       0.057       0.082
host_acceptance_rate           2.7683      0.242     11.427      0.000       2.294       3.243
host_v_email                  -1.9317      0.196     -9.856      0.000      -2.316      -1.548
neighborhood_overview_exist    1.1165      0.144      7.736      0.000       0.834       1.399
host_identity_verified        -0.8830      0.142     -6.217      0.000      -1.161      -0.605
response_time_a_day           -1.5392      0.300     -5.124      0.000      -2.128      -0.950
response_time_a_few_days      -1.9418      0.597     -3.251      0.001      -3.112      -0.771
beds                          -0.1859      0.050     -3.706      0.000      -0.284      -0.088
instant_bookable              -0.5843      0.147     -3.976      0.000      -0.872      -0.296
host_about_exist              -0.2795      0.136     -2.050      0.040      -0.547      -0.012
===============================================================================================
```

Based on the above, we can see that the number of available amenities and the hosts' acceptance rate as well as the existence of neighbourhood overview have a positive impact on becoming "Superhosts". On the other hand, a slower response time by hosts will negatively impact the chance of hosts becoming "Superhosts".

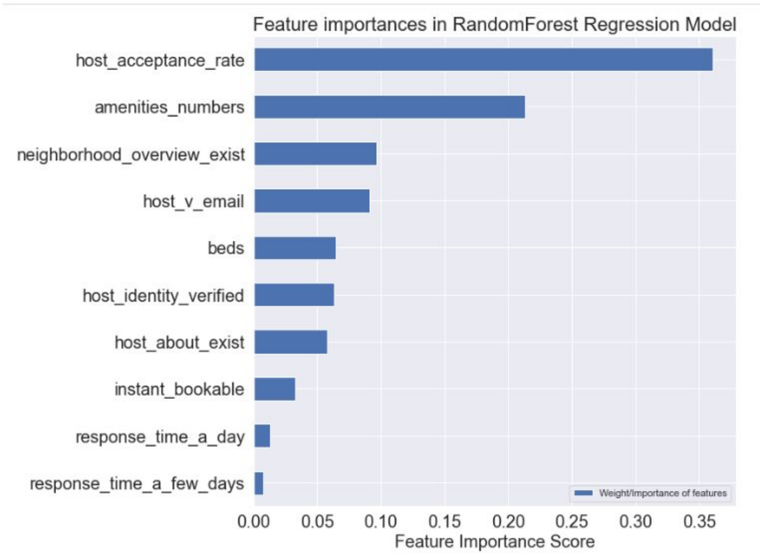
Surprisingly, whether a listing is bookable instantly appears to have a negative impact for listings in Hong Kong, suggesting that this may not be an important feature to guests looking for a place in Hong Kong.

Detailed recommendations to hosts based on the results above are available at the "Recommendation" section below.

In passing, we have also conducted classifications in respect of the above. Details are available at the Appendix of this report for reference.

Model 2: Random Forest Regression

We ran Random Forest Regression on the variables identified for the final Model 1 above to verify the importance of each variable from another perspective. The results are as follows:

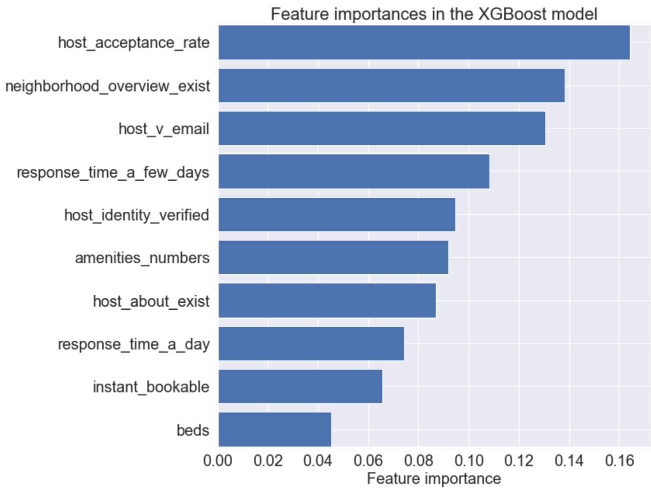


Based on the above, we can conclude that both models produce similar results in that the number of available amenities and the hosts' acceptance rate as well as the existence of neighbourhood overview have the highest impact on becoming "Superhosts".

Model 3: XGBoost

Apart from Random Forest Model, we also built a XGBoost model, an implementation of gradient boost decision trees to identify the usefulness of each feature in the model-building process, to verify the importance of each variable identified in the final Model 1 above.

The results are as follows:



Based on the above, we can conclude that hosts having verification emails and slow response time have the highest importance score. However, we must be cautious that this model does not tell whether a feature positively or negatively impact the prospective of becoming "Superhosts". Indeed, these two factors register high negative impacts on becoming "Superhosts" in the final Model 1 above.

Model 4: Logistic Regression in respect of Amenities

As demonstrated above, the number of available amenities in each listing is the most significant factor determining whether a host can become a "Superhost". Accordingly, we sought to run another logistic regression model focusing on the amenities that are generally available in listings in Hong Kong. In particular, we placed our emphasis on the 35 most common amenities available in listings in Hong Kong.

In arriving at our final logistic regression model in respect of amenities, we have also undertaken a model reduction exercise using the AIC Forward Selection method to further narrow down the number of variables to be used in the model, followed by conducting a (VIF) analysis to identify any existence of multicollinearity issues. That said, certain amenities are almost deemed as necessity as they appear in most, if not all, of the listings which create a multicollinearity problem. As demonstrated below, amenities such as allowing long term stays and air conditioning all have a high VIF score of over 10.

	variables	VIF
0	Shampoo	6.252371
1	Iron	2.925847
2	Hot_water_kettle	1.430071
3	First_aid_kit	1.788546
4	Elevator	2.708379
5	Coffee_maker	1.603831
6	TV	3.687530
7	Cable_TV	1.756567
8	Dedicated_workspace	1.419282
9	Kitchen	4.788073
10	Dryer	1.460525
11	Dishes_and_silverware	3.743478
12	Hot_water	3.334319
13	Extra_pillows_and_blankets	1.721110
14	Fire_extinguisher	2.999285
15	Cooking_basics	2.539328
16	Hair_dryer	6.879310
17	Refrigerator	3.522900
18	Air_conditioning	14.585210
19	Essentials	6.893271
20	Hangers	4.298193
21	Long_term_stays_allowed	15.029286
22	Luggage_dropoff_allowed	1.835552
23	Carbon_monoxide_alarm	1.640830
24	Lock_on_bedroom_door	1.853370

The results of our final logistic regression model in respect of amenities are as follows:

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:      host_is_superhost   No. Observations:                 5056
Model:                            GLM   Df Residuals:                     5031
Model Family:                Binomial   Df Model:                           24
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1256.7
Date:                Mon, 12 Dec 2022   Deviance:                       2513.4
Time:                        19:37:25   Pearson chi2:                 5.36e+03
No. Iterations:                     7
Covariance Type:            nonrobust
==============================================================================================
                                 coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                     -4.5907      0.346    -13.260      0.000      -5.269      -3.912
Shampoo                        1.2221      0.179      6.831      0.000       0.871       1.573
Iron                           0.8694      0.131      6.656      0.000       0.613       1.125
Hot_water_kettle               0.4263      0.169      2.519      0.012       0.095       0.758
First_aid_kit                  0.5245      0.121      4.329      0.000       0.287       0.762
Elevator                      -1.0319      0.115     -8.942      0.000      -1.258      -0.806
Coffee_maker                   0.8132      0.155      5.233      0.000       0.509       1.118
TV                             0.7333      0.152      4.825      0.000       0.435       1.031
Cable_TV                       0.7780      0.188      4.148      0.000       0.410       1.146
Dedicated_workspace            0.5728      0.129      4.447      0.000       0.320       0.825
Kitchen                       -0.8450      0.149     -5.684      0.000      -1.136      -0.554
Dryer                          0.2279      0.123      1.855      0.064      -0.013       0.469
Dishes_and_silverware          1.0339      0.202      5.108      0.000       0.637       1.431
Hot_water                     -0.4396      0.149     -2.940      0.003      -0.733      -0.147
Extra_pillows_and_blankets    -0.3403      0.152     -2.237      0.025      -0.639      -0.042
Fire_extinguisher              0.4785      0.130      3.681      0.000       0.224       0.733
Cooking_basics                 0.3323      0.154      2.159      0.031       0.031       0.634
Hair_dryer                     0.8791      0.207      4.251      0.000       0.474       1.285
Refrigerator                  -0.5541      0.168     -3.303      0.001      -0.883      -0.225
Essentials                    -0.2636      0.203     -1.299      0.194      -0.661       0.134
Hangers                        0.1213      0.159      0.761      0.447      -0.191       0.434
Long_term_stays_allowed        0.2415      0.275      0.879      0.379      -0.297       0.780
Luggage_dropoff_allowed       -0.0908      0.141     -0.643      0.520      -0.368       0.186
Carbon_monoxide_alarm          0.3394      0.130      2.607      0.009       0.084       0.594
Lock_on_bedroom_door          -0.1268      0.129     -0.980      0.327      -0.380       0.127
==============================================================================================
```

Based on the above, we can conclude that the availability of the following amenities has a positive impact towards becoming "Superhosts": (i) shampoo, (ii) iron, (iii) hot water kettle, (iv) first-aid kit, (v) coffee maker, (vi) television, (vii) cable television, (viii) dedicated workspace, (ix) dryer, (x) dishes and silverware, (xi) fire extinguisher, (xii) basic cooking utensils, (xiii) hair dryer,

(xiv) hangers, (xv) allowing long term stays, (xvi) carbon monoxide alarm and (xvii) locks for bedrooms' doors.

Specifically, (i) shampoo, (ii) iron and (iii) television are the top-three amenities that have the strongest impacts towards becoming "Superhosts".

Conversely and surprisingly, having (i) kitchen and (ii) elevator in the listing negatively impacts a host becoming "Superhosts" significantly. Because of such peculiar conclusions from the results above, we dug further into the original data sets.

In respect of having kitchen in the listing, we found that one of the possible reasons might be that hosts may only list kitchen as an amenity when the property type is a room, as opposed to an entire flat. Through our spot checking on some properties on Airbnb Hong Kong, we also found that some listings have highlighted the kitchen as an amenity, but do not offer additional amenities such as cutleries, utensils, cooking equipment, etc, or have faulty facilities which takes away from the staying experience.

In respect of having elevators in the listing, we found that in Hong Kong, many listings have elevators purely because the listing is based around high density locations such as Mong Kok, Tsim Sha Tsui and Yau Ma Tei. As such, these properties are usually apartments which require elevators in order to go up and down. That said, since it is something that is commonly highlighted in the majority of listings, we have also noticed a trend that "Superhosts" tend to not list it within their amenities, resulting in a negative coefficient and impact to the "Superhost" response variable.

Some other common amenities you would expect to exist in listings in Hong Kong such as hot water and essentials also have a negative impact on becoming a "Superhost". While having more amenities highlighted in the listing does help contribute towards becoming a "Superhost", we also noticed that some hosts tend to just highlight any amenities which could potentially be applied to their listing, which dilutes the quality of amenities analysis that can be done. As such, we do see some standard amenities which have a negative impact on becoming a "Superhost".

Recommendations and Conclusions

Based on the results from various models above, we can conclude that the number of available amenities and the hosts' acceptance rate as well as the existence of neighbourhood overview have a positive impact on becoming "Superhosts". Meanwhile, a slower response time by hosts will negatively impact the chance of hosts becoming "Superhosts".

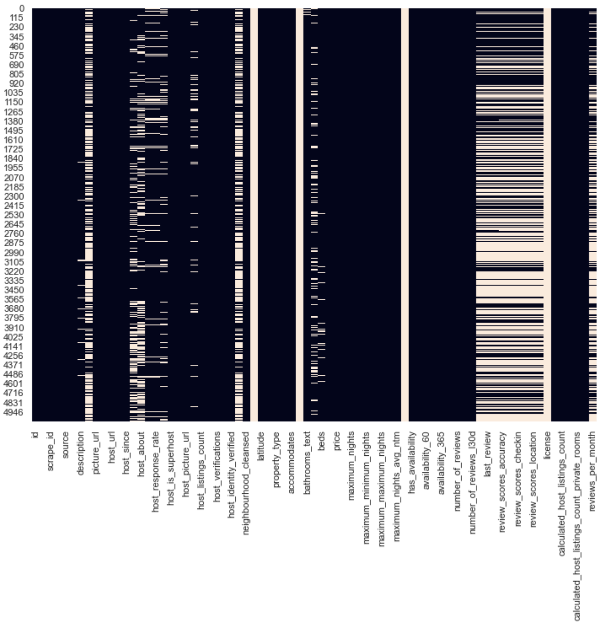
In an attempt to improve its earnings, Airbnb should urge and provide further incentives to all hosts to adopt the following practices in a bid to achieve "Superhost" status:

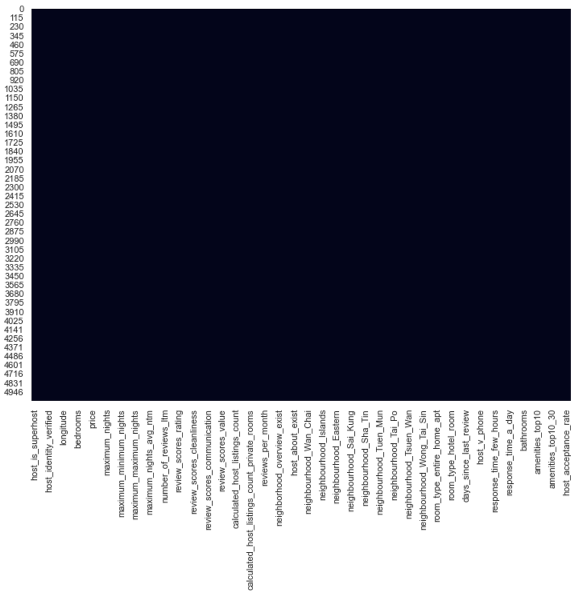
1. Hosts should strive to reduce their response time to within a day;
2. Hosts should accept every booking unless there are any sounding reasons or extenuating circumstances. To the extent that hosts are not prepared to accept bookings for a certain period of time, hosts should remove their listings for such periods to maintain a high booking acceptance rate;
3. Hosts should always include an overview of the neighbourhood of the listing. Providing such background information will enhance bookings, to say the least; and
4. While highlighting amenities does assist in helping a host become a "Superhost", hosts should be cautious on just applying any amenities and should give thought as to how much added value their amenities give to guests. As such, hosts should selectively invest in amenities which add the most value to guests. Aside from the basic amenities that most listings should have, i.e. Wifi and air conditioning, our observations from the logistic regression shows that there are certain amenities which show a stronger positive impact for hosts to become a "Superhost". For example, having a fully equipped kitchen with items such as hot water kettle, coffee maker, dishes and silverware, and basic cooking utensils. Hosts should also have added amenities in the bathrooms such as shampoo and hair dryers which add to the overall experience. Within bedrooms, having amenities such as hangers, and locks on the bedroom doors. Other general amenities which also added value to being a "Superhost" include having televisions (preferably with cable), dedicated workspace, safety equipment (i.e. fire extinguisher and first-aid kit) and flexibility for longer term stays.

Appendix

Comparison of missing data before and after cleaning

The graphs below demonstrate that there are no further missing data after our robust data cleaning processes.



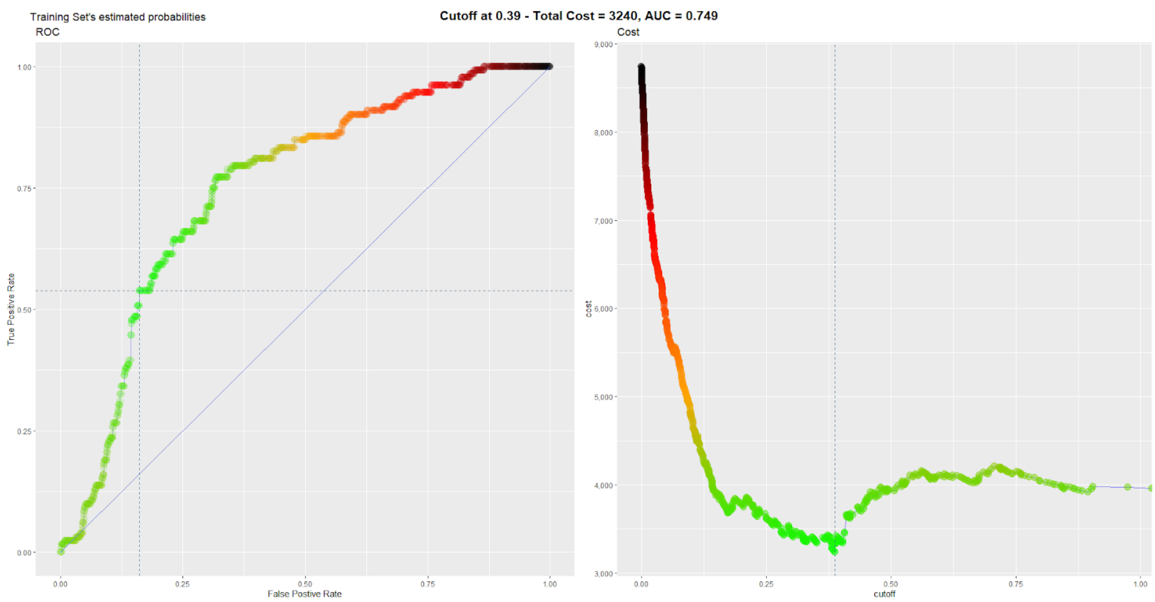


(Before) (After)

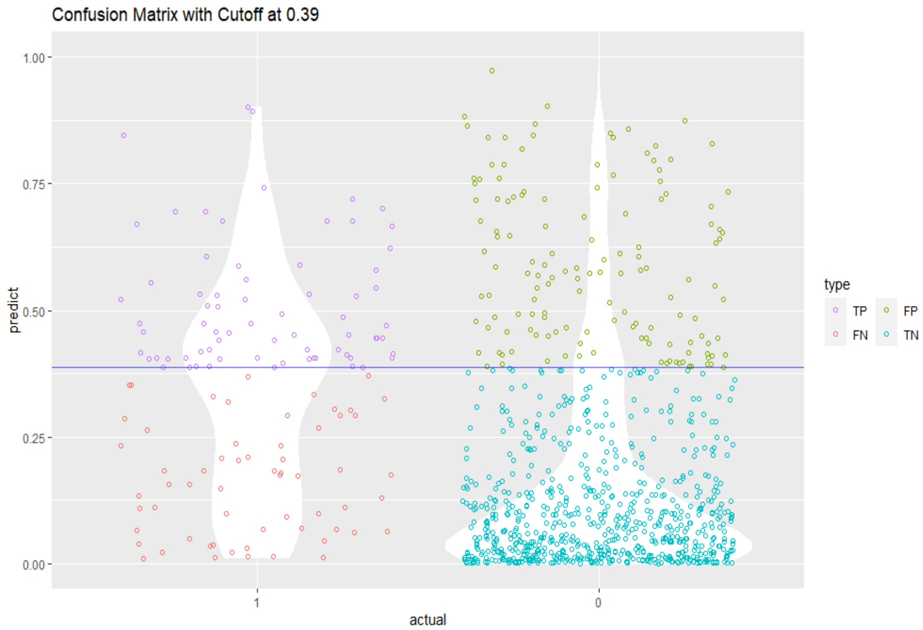
Cost-sensitive classification & error measurements in respect of Logistic Regression

[1] For Model 1: Logistic Regression on Feature Data

Based on the 5-folded cross-validation approach together with an assumption that the cost of wrongly predicting the hosts to become a Superhost is triple the cost of not predicting some Superhost in advance, we determine the optimal cut-off point to be 0.39, which could minimise the total cost. That will lead to an AUC of 0.749.



Based on the optimal cut-off level, we get the confusion table and the confusion matrix as below:



	FALSE	TRUE
0	733	141
1	65	67

[1] The sensitivity is 0.5379

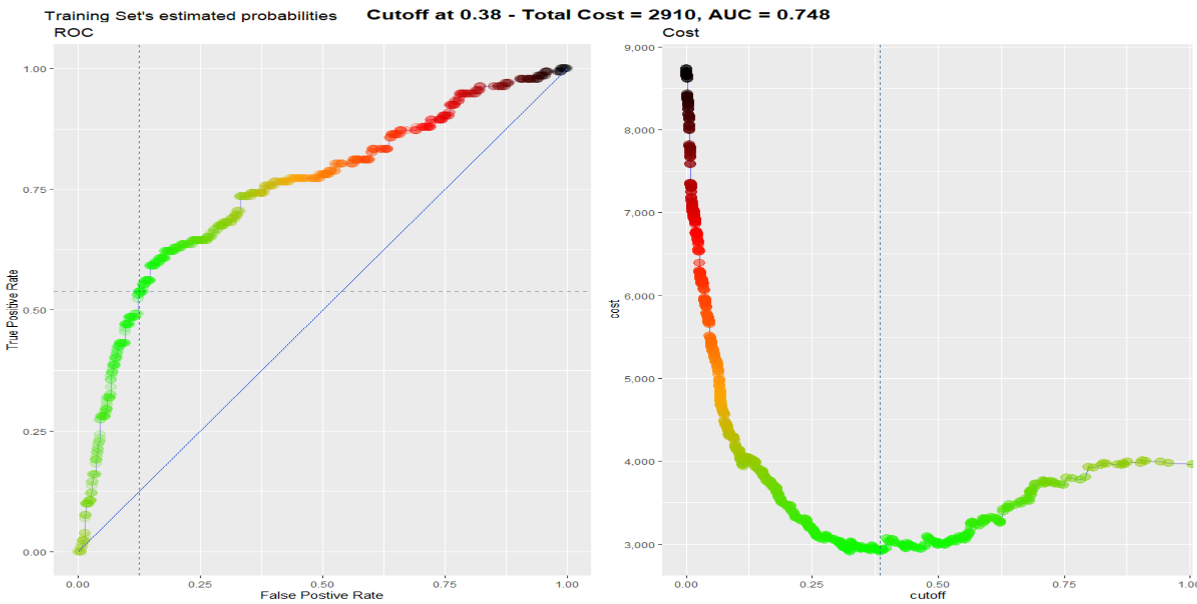
[2] The specificity is 0.8387

[3] The Misclassification Rate is 0.2008

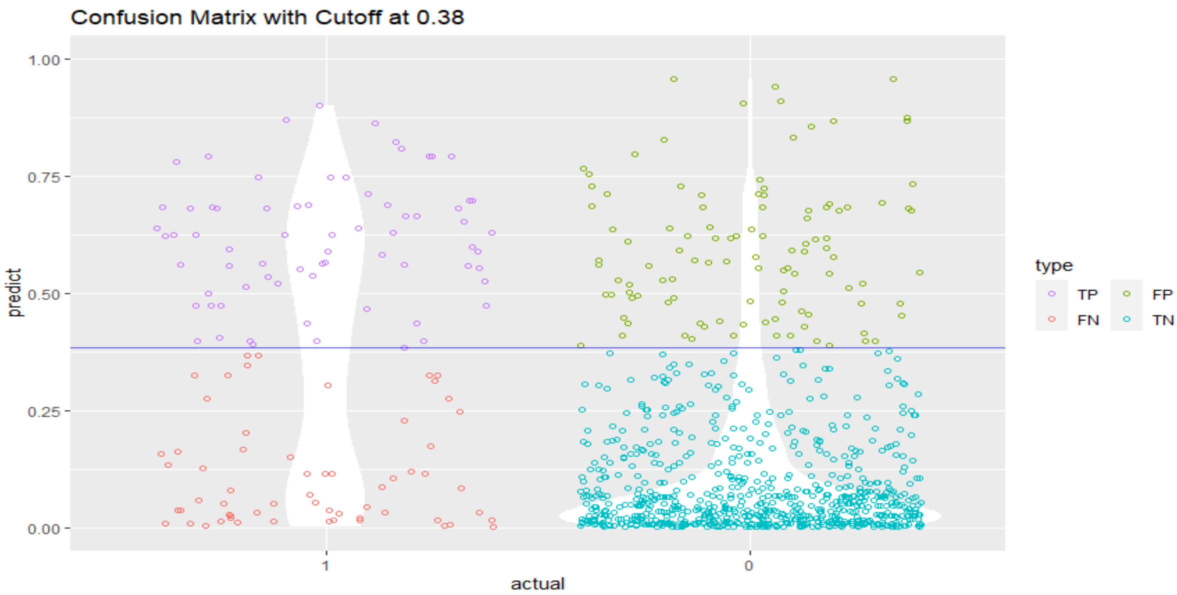
It could be concluded that the model gives reasonable prediction accuracy with the TP rate (Sensitivity) of 53.79%, whereas the overall Misclassification Rate is 20.08%.

[2] For Model 2: Logistic Regression on Amenity data

Based on the 5-folded cross-validation approach together with an assumption that the cost of wrongly predicting the hosts to become a Superhost is triple the cost of not predicting some Superhost in advance, we determine the optimal cut-off point to be 0.38, which could minimise the total cost. That will lead to an AUC of 0.748.



Based on the optimal cut-off level, we get the confusion table and the confusion matrix as below:



	FALSE	TRUE
0	766	108
1	62	70

[1] The sensitivity is 0.5379

[2] The specificity is 0.8764

[3] The Misclassification Rate is 0.168

It could be concluded that the model gives reasonable prediction accuracy with the TP rate (Sensitivity) of 53.79%, whereas the overall Misclassification Rate is 16.8%.

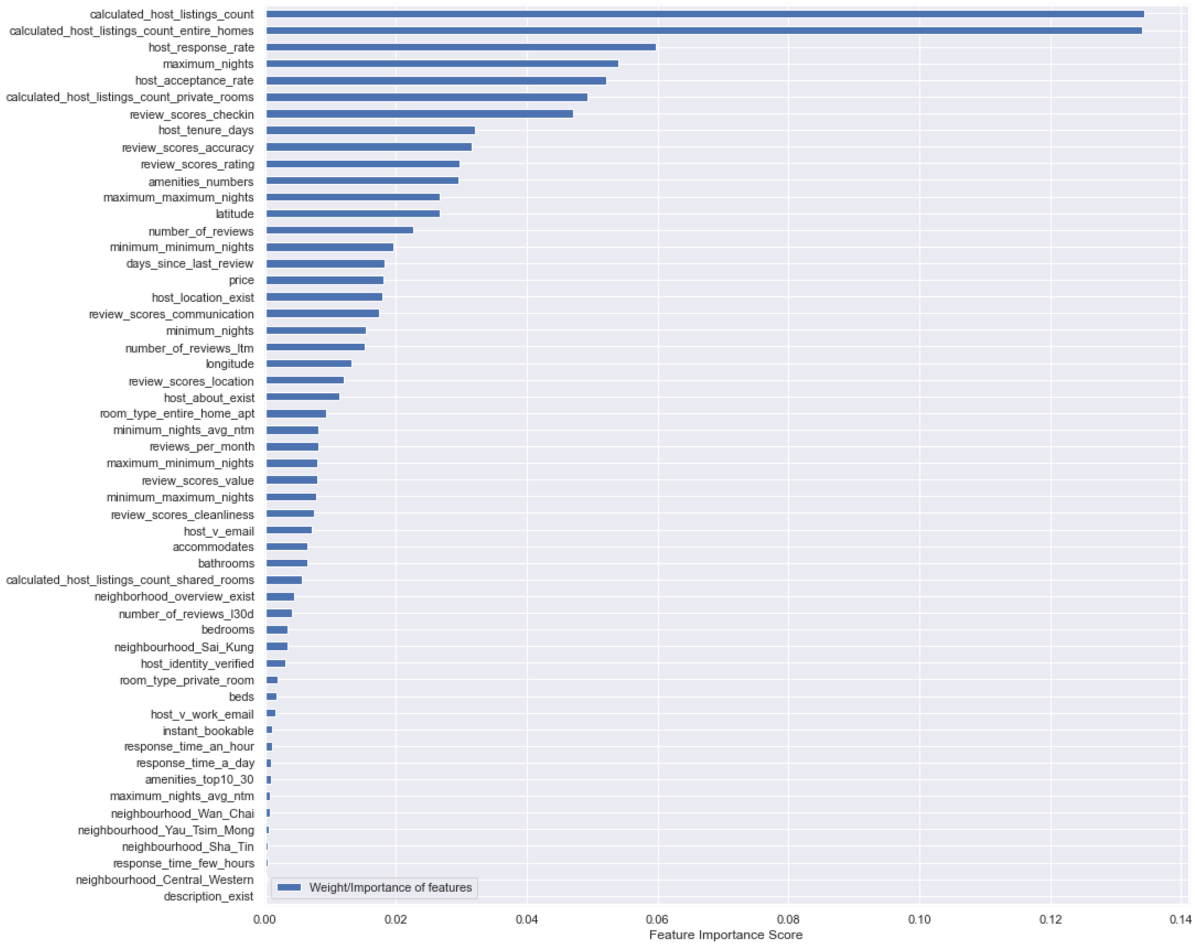
<u>Using Random Forest Regression Model and XGBoost Regression Model to verify the results in Logistic Regression Models</u>

We have employed the popular Random Forest Regression Model and XGBoost Regression Model to compare the results run through Logistic Regression Model.

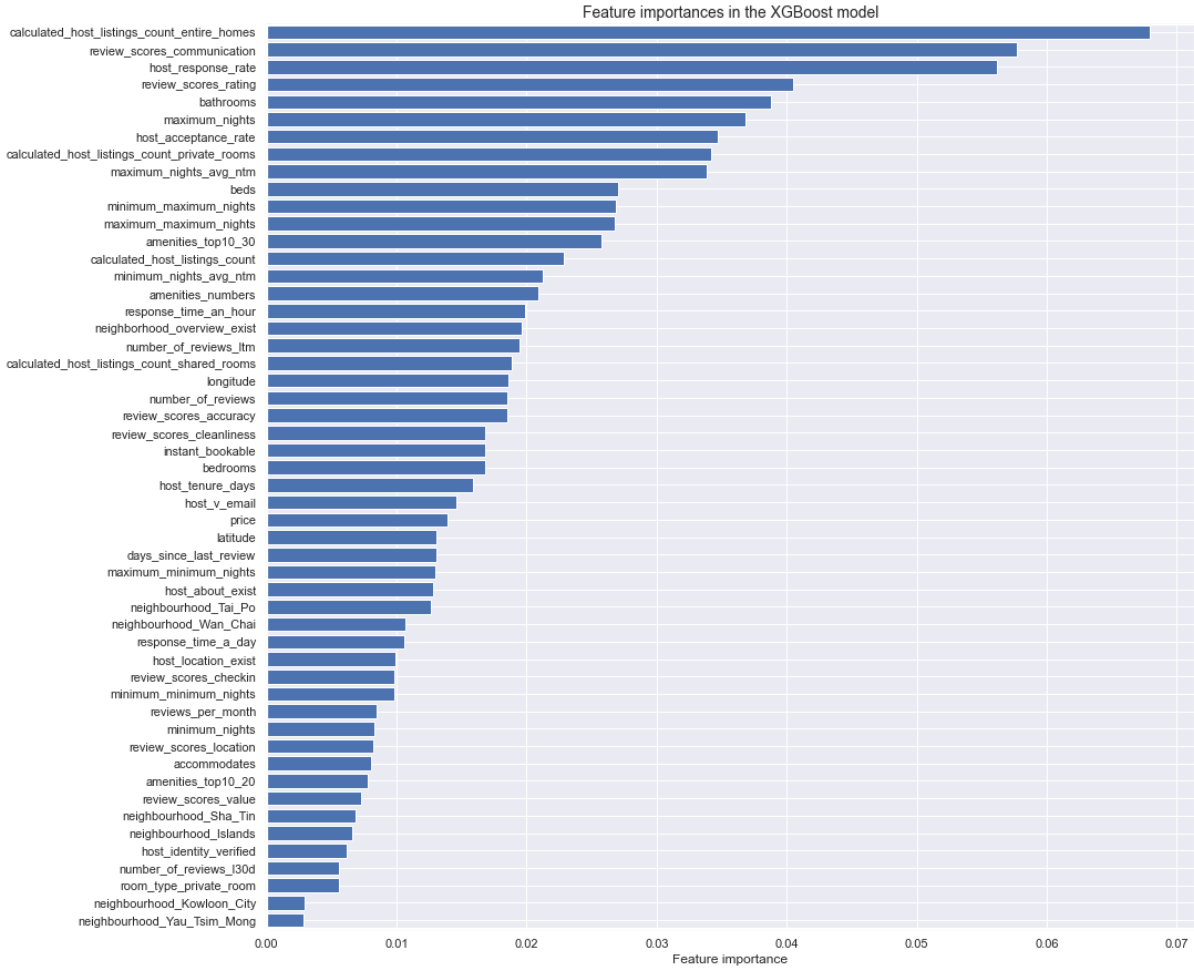
The results of these models will generate scores indicating how useful or valuable each feature is in the construction of the boosted decision trees within the model. The more a variable is used to make key decisions with decision trees, the higher its relative importance. As such, feature importance can be used to interpret our data to understand the most important features that define our predictions.

Below are results from running these two models with all variables included as an initial guidance for our further analyses:

<u>Random Forest Regression Model</u>



XGBoost Regression Model



Model evaluation metrics: MSE and R^2 for the models adopted for our feature analyses

```
Goodness Fit on the Models (Train/Test Split) with all cleaned variables:

Performance Metrics for Test Set
--------------------------------


Model 1: Logistic Regression on Feature Data (MSE): 0.11076
Model 1: Logistic Regression on Feature Data (R^2): 0.02835

Model 3: RandomForest Classification on Feature Data (MSE): 0.06917
Model 3: RandomForest Classification on Feature Data (R^2): 0.31843

Model 4: XGBoost Classification on Feature Data (MSE): 0.05237
Model 4: XGBoost Classification on Feature Datat (R^2): 0.31843


Performance Metrics for Train Set
-----------------------------------
Model 1: Logistic Regression on Feature Data (R^2): 0.19819
Model 3: RandomForest Classification on Feature Data (R^2): 0.54083
Model 4: XGBoost Classification on Feature Data (R^2): 0.76639
```

References

Websites

Airbnb, 'About Superhosts' <https://www.airbnb.com/help/article/828> accessed 10 December 2022.

Airbnb, 'Airbnb Service Fees' <https://www.airbnb.com/help/article/1857> accessed 11 December 2022.

Airbnb, 'How to become a Superhost' <https://www.airbnb.com/help/article/829#section-heading-2-0> accessed 10 December 2022.

Airbnb, 'Superhost' <https://blog.atairbnb.com/superhost/> accessed 10 December 2022.

AirGMs Technologies Inc., 'What Is Airbnb Superhost Status and Is It Worth Getting?' <https://www.igms.com/airbnb-superhost/> accessed 10 December 2022.

Inside Airbnb, 'Hong Kong' <http://insideairbnb.com/get-the-data/> accessed 10 December 2022.

Maria Pengue, 'Airbnb Statistics: Demographics, Superhosts, Airbnb Plus in 2022" <https://writersblocklive.com/blog/airbnb-statistics/> accessed 10 December 2022.

Wall Street Journal, 'Airbnb Inc. CI A' <https://www.wsj.com/market-data/quotes/ABNB> accessed 10 December 2022.

---

[i] Wall Street Journal, 'Airbnb Inc. CI A' <https://www.wsj.com/market-data/quotes/ABNB> accessed 10 December 2022.

[ii] Airbnb, 'Superhost' <https://blog.atairbnb.com/superhost/> accessed 10 December 2022.

[iii] Airbnb, 'About Superhosts' <https://www.airbnb.com/help/article/828> accessed 10 December 2022.

[iv] iBid.

[v] Airbnb, 'How to become a Superhost' <https://www.airbnb.com/help/article/829#section-heading-2-0> accessed 10 December 2022.

[vi] iBid.

[vii] AirGMs Technologies Inc., 'What Is Airbnb Superhost Status and Is It Worth Getting?' <https://www.igms.com/airbnb-superhost/> accessed 10 December 2022.

[viii] Airbnb, 'Airbnb Service Fees' <https://www.airbnb.com/help/article/1857> accessed 11 December 2022.

[ix] Maria Pengue, 'Airbnb Statistics: Demographics, Superhosts, Airbnb Plus in 2022" <https://writersblocklive.com/blog/airbnb-statistics/> accessed 10 December 2022.

[x] Inside Airbnb, 'Hong Kong' <http://insideairbnb.com/get-the-data/> accessed 10 December 2022.

[xi] Within the dataset available at Inside Airbnb, 531 out of 5056 data entries are listings made by "Superhosts", resulting in 10.5% of hosts being "Superhosts" in Hong Kong.