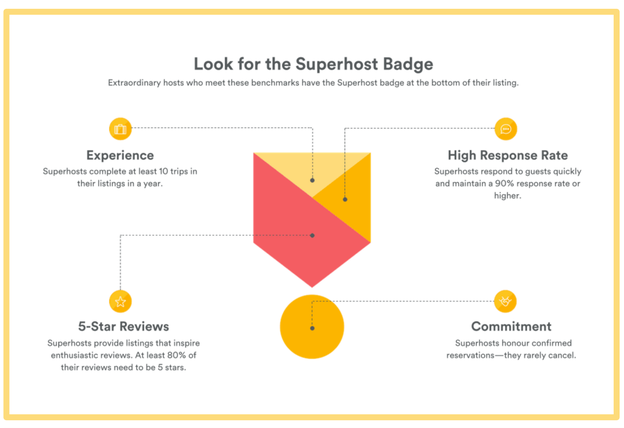
Predicting Superhost Status

Shirley Xu & Tiffany Xiao

Introduction

With over 150 million[1] users, 640,000[1] hosts and 4 million[1] listings, there is no question that Airbnb and its hosting service is popular in today's world. The popularity of this hosting service brings to mind the question "what qualities makes a person a good host?"

Airbnb has a term for exceptional hosts - "Superhosts". Superhosts are described by Airbnb as "experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests"[2]. However, this is a vague description that does not tell users about why they should look for hosts with Superhost status or how they can become a Superhost. Thus, our project focuses on answering the question "**How can we best predict superhost status for an Airbnb host?**"



Predictors (categorical, numerical)

- `host_response_time` – amount of time it takes a host to respond (description based)
- `host_response_rate` - percentage based on how often a host responds to inquiries
- `host_neighbourhood` – neighbourhood the host is from
- `host_total_listings_count` – number of total listings the host has
- `num_verification_methods` – number of ways of verification the host has activated
- `host_identity_verified` – binary variable indicating whether or not a host's identity has been verified
- `number_of_reviews` – number of reviews the host has
- `host_has_profile_pic` – binary variable indicating whether or not a host has a profile picture
- `review_scores_rating` – percentage based on the hosts' review score
- `host_about_yesNo` – binary variable indicating whether or not a host has a self-introduction online
- `host_days` – number of days the user has been a host

Data & Methods

In order to answer our question, we analyzed available information about Airbnb listings in Boston. We got the listings data for Boston from Inside Airbnb[3], an independent, non-commercial set of tools and data for people to better understand how Airbnb is really being used.

The listings data contains 4870 observations – each observation is a listing with values for 96 variables such as `id` (a unique id for each listing), `listing_url` (the website link for the listing), and more.

Note: Since our project is focused on predicting Superhost status, we removed duplicate observations (multiple listings that belong to the same host), leaving 2705 observations. After further cleaning, we were left with 1735 observations.

Of the 96 variables we selected 11 predictors that seemed to be related to host information and guest reviews to predict which hosts are Superhosts. Our binary response variable is `host_is_superhost` where "f" means false (the host is not a Superhost) and "t" means true (the host is a Superhost).

After data cleaning and pre-processing, we applied two techniques to create our model – **logistic regression** and **classification tree**.

Results

Logistic Regression

We built two logistic regression models – one with an unbalanced dataset and one with a balanced dataset. Our first logistic regression model (unbalanced) correctly predicts whether a host is a Superhost 79.26% of the time on the test set. After generating a confusion matrix, we discovered that our model yields a test with a high sensitivity but low specificity (sensitivity: 0.9346, specificity: 0.3894). We also created an ROC curve, and the area under the ROC curve is 0.8181843, which means it is a good model according to the guide for classifying the accuracy of a diagnostic test[4].

Note: Our unbalanced dataset was in favor of host_is_superhost == 'f' with only 27.9% of the hosts being Superhosts.

Our second logistic regression model (balanced) correctly predicts whether or not a host is a Superhost 70.66% of the time on the test set. The confusion matrix shows us that our model yields a test with similar sensitivity and specificity values (sensitivity: 0.6748, specificity: 0.7395).
The results from the logistic regression models suggests that the performance of a logistic regression model depends on how balanced the dataset is.

Classification Tree

Then, we decided to build a classification tree to see if the variables used in the tree construction are the same as the significant predictors in our logistic regression models. The classification tree correctly predicts whether a host is a Superhost 83.87% of the time on the test set.

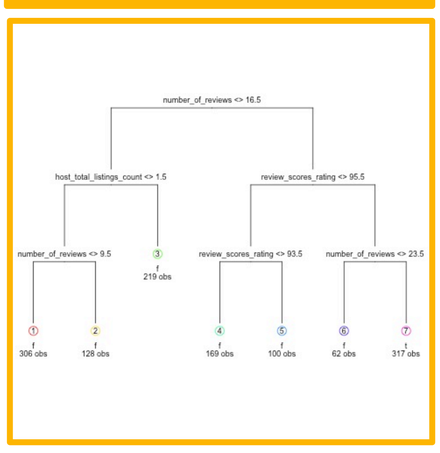
Significant Predictors

Logistic Regression: `host_response_rate`, `host_identity_verified`, `number_of_reviews`, `review_scores_rating`, `num_verification_methods`, `host_about_yesNo`

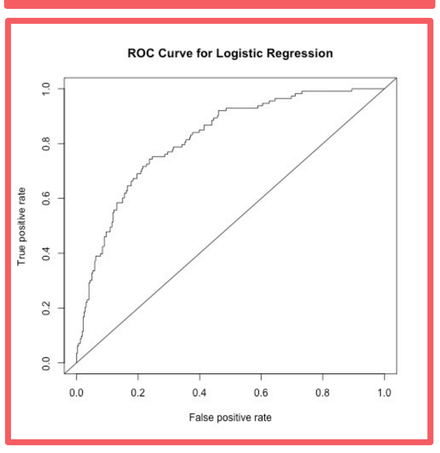
Note: `num_verification_methods` and `host_about_yesNo` are not significant in our second logistic regression model

Classification Tree: `number_of_reviews`, `host_total_listings_count`, `review_scores_rating`

Classification Tree



ROC Curve



World Cloud

Conclusions & Future Work

Both techniques identified `number_of_reviews` and `review_scores_rating` as significant predictors when identifying whether or not a host is a Superhost. Thus, the most reliable way to predict Superhost status is most likely by finding out how many 5-star reviews a host has.

For future work, we would like to research more on how to impute missing values for some of our numeric variables. We would also like to attempt to apply a random forest model on our dataset.

References

[1]Smith, Craig. "100 Amazing Airbnb Statistics." DMR, 5 Dec. 2017, expandedramblings.com/index.php/airbnb-statistics/.
[2]"What Is a Superhost?" What Is a Superhost? | Airbnb Help Center, www.airbnb.com/help/article/828/what-is-a-superhost.
[3]"Inside Airbnb. Adding Data to the Debate." Inside Airbnb, insideairbnb.com/get-the-data.html.
[4]Tape, Tom. The Area Under an ROC Curve, gim.unmc.edu/dxtests/roc3.htm.

Acknowledgements

This project was completed in partial fulfillment of the requirements of SDS293: Machine Learning. This course is offered by the Statistical and Data Sciences Program at Smith College, and was taught by R. Jordan Crouser during Fall 2017.