Paper 4055-2018

Airbnb New York City: Demystifying the Superhost Program

Man Singh, Soumya Kar Choudhury, Rohit Banerjee and
Andres Manniste, Oklahoma State University

ABSTRACT

This paper investigates the relative importance of the qualifying factors stated by Airbnb to promote their Superhost program and makes an effort to discover other implicit factors like property name, description, amenities, house rules that play a significant role in awarding the Superhost status. The goal of this project is to help hosts improve the experience of their guests and provide them with a framework that will guide them to increase their chances in the 'Superhost' program. The study utilizes the listings from New York City to predict the probability of becoming a Superhost based on text mining. Various statistical models such as Logistic Regression, LARS, SVM and Decision Tree were built to achieve this objective. Decision Tree was the best model with a misclassification rate of 12% and sensitivity of 52%. Another set of linear and non-linear regression models were built to predict the listing price of all Airbnb listings in New York City and in this case, stepwise regression model outperformed the other alternatives with an R Squared value of 55%.

INTRODUCTION

Airbnb is a peer to peer platform that serves as an online broker between property owners willing to allow short term rentals of their properties to travelers. Hosts with creditable service are rewarded with a Superhost status as a recognition; this could result in an improved host-guest experience/relationship cycle. Customers looking for places pay attention to this and gravitate towards renting from hosts who have earned the Superhost badge of approval. This is a formula that could keep both Superhosts and customers happy. Airbnb has published factors used as qualifying criteria in this process:

- Overall ratings of 4.8+
- 90% response rate within 24 hours
- Host at least 10 stays a year
- Honor confirmed reservations with no cancellations
- Maintain at least a 50% review rate

PROBLEM STATEMENT

There appears to be other factors that play important roles in selection of the "Superhost", especially how a listing is described on the Airbnb website using fields such as name of the property, description, amenities provided and house rules; Airbnb corporate appears to include such factors for its Superhost recognition.

To validate the aforementioned hypothesis, different modeling techniques such as Text Mining, Logistic Regression, LARS and Gradient Boosting have been utilized. Furthermore, customers' reviews have also been analyzed to explore if reviews for Superhost are different compared to a non Superhost and if differences exist, what impact that has on being recognized as a Superhost. In addition to Superhost prediction, a secondary model was also built to predict the listing price and analyze the variables affecting price.

DATA

Airbnb data is publicly available on the Inside Airbnb1 website. The data collected from this source was a demographic listing file, outlining all Airbnb properties in New York City as well as another review file providing customer reviews in between 10/13/2008 and 9/8/2018. The reason the city of New York was

[1] http://insideairbnb.com/get-the-data.html

selected is based upon its popularity as a tourist destination as well as the city being the financial hotspot of the US. For scoring purposes, Airbnb data for Boston city was collected from the same website.

EXPLORATORY DATA ANALYSIS

An exploratory data analysis was conducted using SAS® Enterprise Guide to observe the distribution of the numeric variables in the *listings* dataset. The *listings* file provides all the property information in New York City such as name, description, amenities, house rules, host verification, neighborhood, number of reviews and review ratings, bedrooms, bathrooms and any other relevant information about a house. The overall dataset had 96 different variables and 50,220 listings among which 7,905 were of Superhosts.

It was evident from the exploratory analysis result that several variables were heavily right skewed. There are multiple variables such as maximum nights and minimum nights with lot of outliers. Table 7 highlights some of the basic information about the most relevant numeric variables used in the analysis.

DATA PREPARATION

Superhost status is awarded on a quarterly basis and the data files used in the analysis are as of 9/8/2018. Therefore, the *listings* file utilized in this analysis contains Superhosts and non Superhosts as of the second quarter of 2018. To maintain the data consistency, the same time period has been chosen for customer reviews details from the *reviews* file.

As part of data cleaning, the variables have been selected on the basis of data consistency, and relevancy towards the goal. For example, the dataset had variables with only one level or variables like zip code and physical coordinates which weren't required for the analysis. To further prepare and analyze fields in the *listings* file, text analytics was performed on the text fields in the *listings* file to extract relevant text topics. Through this analysis, the significant text variables were determined to be Description, Name, Amenities and House Rules. The *reviews* file was joined through listing number to the *listings* file in order to separate out reviews by Superhosts and non-Superhosts; this was necessary to evaluate varying sentiments in both cases.

Text variables neighborhood overview and interaction were converted into binary flags, depending upon the presence of text in that particular variable. The above mentioned attributes provide customers with information on whether the host property provides a pleasant view of the neighborhood and what the different methods of interactions with the host, respectively. The text field host verification contained several methods of verification enclosed in a string. Each verification method was parsed out and converted into binary variables. The most and least common methods of verification were not considered as those methods did not provide significant lift to the model. A new feature called days since last review was created that acts similar to a "recency" measure and it holds data about the days since the listing received its last review, as of 9/8/2018 and was added to the final dataset for modeling.

MODELING

Methodology for Superhost Prediction

1. Models were built separately on the text variables, numerical variables as well as a combination of both. The goal was to include all the relevant variables in the *listings* files to make a model that predicts Superhosts, apart from the ones already mentioned by Airbnb in their qualifying criteria.
2. For each approach mentioned above, four modeling techniques were used: Support Vector Machine, LARS, Regression and Decision tree. The choice of modeling techniques was based on the fact that all the four models are able to handle higher dimensionality with relative ease or are easy to interpret.
3. Data replacement, data filtering, text modeling, transformation and imputation were performed on SAS® Enterprise Miner™.
4. Data filtering was heuristic and used to exclude extreme values of all numeric variables. Obvious outliers (such as a 10-digit number of bathrooms) were removed from the modeling dataset. Imputation was done by the mean method for all the interval variables while categorical variables

were grouped to reduce the number of levels. A log transformation was performed on the interval variables.

5. Binary variables such as the target (host_is_superhost) were converted into a numeric format (0 or 1) from categorical (f or t). Variables with multiple levels such bed type and property type were grouped by common sense to have two overall levels, for the ease of using them in the model.
6. The data had 4 important text variables Name, Description, House rules and Amenities as mentioned under Data Preparation section. Text topics were created for each of these variables and used as input into the models. The decision of the apt amount of text topics was based on the frequency of parsed single terms and their weighted importance.
7. The text topics with their terms can be found in Table 2, 3, 4 and 5 in the appendix.
8. All the exported data from the text topics were merged and a final dataset with all the text topics and other numeric variables was used for modeling.
9. The Decision tree model in both the numeric and text approach was selected as the best model, selected on the basis of sensitivity, misclassification rate and relevance for the business problem. Table 1 shows the performance metrics for all models. Table 1 also shows that the models with only text topics showed very low relative sensitivity hence this approach of modeling was not pursued further.

Model	Validation Sensitivity	Validation Misclassification Rate
Decision tree(non-text)	59%	11.38%
Decision tree(all variables)	51%	11.58%
SVM (all variables)	43%	11.96%
Regression(non-text)	42%	12.05%
SVM(non-text)	42%	12.07%
Regression (all variables)	39%	12.41%
LARS(all variables)	31%	14.04%
LARS(non-text)	28%	14.67%
Decision tree(text only)	5%	15.59%
Regression(text only)	10%	15.69%

Table 1: Model performance Sensitivity and Misclassification Rate

Methodology for Price Prediction

The secondary objective of this paper was to determine the factors affecting the listing price based on the currently available data.

1. Four regression models were built namely Stepwise, Backward, Forward and LARS.

2. Stepwise regression with an adjusted R Square of 55% was selected as the best model. The same data preparation and manipulation methods were applied to the dataset used for predicting price.

RESULTS

Results for Superhost Prediction

1. The Decision tree model was used as the final predictive model for data scoring. LARS and regression results were also considered to have a better explanation of the results. Table 8 in the appendix shows the variable importance table for all the significant variables in the model.
    i. The number of reviews, cleanliness review scores, host response rate and number of host listings were the most important variables in the final model, along with other variables as seen in Table 8.
    ii. An interesting finding from this model was that three text topics from house rules, amenities and listing description were part of the decision tree and the English rules. This

demonstrates how including important terms in the listing of the Airbnb website could enhance the chances of having a Superhost recognition.

iii. If the number of reviews for a host is greater than 32, cleanliness score is greater than 9.5, host response rate is greater than 84%, value score is greater than 9.5 and days since last review is less than 153, then the chance of being recognized as a Superhost is 77%.

iv. If the number of reviews is between 10 and 20, accuracy score is greater than 9.5, number of listings is more than 2, cleanliness score is greater than 9.5, host response rate is greater than 85% and the listing description has terms like “fully”, “equip”, “furnish”, “enjoy” and “offer” then the chance of being Superhost is 78%.

v. If the host response rate is greater than 85%, the amenities has terms like “intercom”, “wireless intercom”, “buzzer”, “TV”, cleanliness scores are almost equal to 10, number of reviews is between 17 and 21 and number of host listings is less than 2, then the chance of being a Superhost is 80%.

2. Results from LARS were used to determine variable importance using the standardized estimate. Table 9 in the appendix shows the details of the variable importance.
   i. Table 9 reiterates the peak importance of the number of reviews and host response rate in Superhost prediction. If the listing is situated in Manhattan borough, then the chances of Superhost becomes less. This corroborates with the fact that the highest number of Superhosts belong to Brooklyn borough (42%).
   ii. A listing which is an entire home/apartment has a negative effect on Superhost recognition. The other room types, private room and shared room provided better prospects towards a Superhost.

3. For quantifying the increase in probability of Superhost predictions based on the significant variables in the model, the odds ratio from the regression model was used. The values of all odds ratios can be found in Table 10 in the appendix.
   i. If the description field in a listing contains terms like “floor”, “house”, “entrance”, “brownstone”, “garden”, then the odds of being a Superhost goes up by 42%.
   ii. If the Airbnb has amenities like “dish”, “refrigerator”, “silverware”, “stove” and “oven” then the odds of being a Superhost goes up by 33%.
   iii. If the Airbnb has amenities like “first aid kit” and “card” then the odds of being a Superhost goes up by 35%.
   iv. If the Airbnb has amenities like “wide access”, “wide doorway”, “wide clearance” then the odds of being a Superhost goes up by 43%.

Results for Price Prediction

Here are the following results from the price prediction model:

1. The variables with a significant positive effect on price were number of bathrooms, location scores, number of bedrooms, cleaning fee, accommodation, Manhattan location (compared to Staten Island) and some interactions between property type and location, room type and location.

2. The variables negatively affecting price were minimum nights, Bronx, Queens and Brooklyn location (compared to Staten Island) and some interactions between property type and location, room type and location. The main and interaction effects along with their estimates can be found in Table 6 in the appendix.

3. There are significant effects on comparing apartments against all other property types in Manhattan, Bronx and Queens. In Bronx and Queens, the nightly price reduces when the property type is an apartment while in Manhattan, the nightly price increases when it’s an apartment. Prices are not significantly affected by property type in Brooklyn.

4. In comparing prices for different room types, there are significant effects across the board for renting an entire home compared to a private room. Price of a room significantly decreases when it is shared with someone versus when it’s a private room. This holds true, though, only for Manhattan. Other neighborhoods are not significantly affected by the room type (private or shared).

Results for Sentiment Analysis

Sentiment Analysis performed on the *reviews* file separately for the current Superhosts and hosts revealed that 4.4% of the reviews were negative for the non-Superhosts while only 0.5% of the reviews were negative for the current Superhosts. These numbers indicate that reviews and review scores play a significant role in naming a Superhost for Airbnb.

SCORING

The *listings* dataset from Boston was used for scoring purposes. The dataset was subject to the same data preparation steps as the dataset from New York City. Creating exactly the same text topics for Boston as for New York was not possible because of the varying textual field differences and the smaller number of Airbnbs in Boston leading to inevitably less text terms.

So, instead of using the Numeric and Text model the Numeric only model (Decision Tree) with the least misclassification rate was used for scoring. From the Boston dataset, 879 cases out of 1377 were correctly predicted by the NYC model.

FUTURE PROSPECTS

Census data can be also leveraged to further enhance the accuracy of the model by including socio-economic and demographic factors. This analysis can be expanded to include other cities in the United States to study the varying factors in determination of a Superhost, depending on the location of the Airbnb.

Time series forecasting techniques can also be performed on price data to better forecast the listing prices based on their historical rates. This would ensure a compliance on any sudden fluctuations in price.

CONCLUSION

Apart from the five criteria prescribed by Airbnb for a Superhost consideration, this paper attempts to expand the horizons of great hosting by use of text analytics and predictive modeling. The study of New York City shows that Airbnb hosts should encourage the visitor to write reviews for every visit. While keeping the property clean is an evident requirement, having varied and useful amenities lends extra credit to a host's portfolio. A precise and descriptive listing also helps boost the chances of gaining a Superhost status. If the Airbnb is located in Manhattan, the nightly price always stays higher than the other neighborhoods with entire homes being costlier than private rooms and other property types.

A combination of insights from all the models were used to determine comprehensive and scalable insights. Among the various models built, decision tree provided the best interpretable results while including most of the relevant variables from the *listings* data.

REFERENCES

Barron Kyle, Kung Edward, Proserpio Davide. October 2017. "The Sharing Economy and Housing Affordibility: Evidence from Airbnb".
https://www.aeaweb.org/conference/2018/preliminary/paper/ykYrh4Gd

Gunter Ulrich, November 2017. "What makes an Airbnb host a Superhost? Empirical evidence from San Francisco and the Bay Area". https://www.journals.elsevier.com/tourism-management

https://learnairbnb.com/earning-a-5-star-review-on-airbnb/ (Accessed October 2018)

CONTACT INFORMATION

Andres Manniste andres.manniste@okstate.edu Man Singh man.singh@okstate.edu

Rohit Banerjee rohit.banerjee@okstate.edu Soumya Ranjan Kar skarcho@okstate.edu

APPENDIX

Topic ID	Doc Cutoff	Term Cutoff	Term	Number of Docs
1	0.32	0.276	"+private +room +private room +bedroom +bath"	7751
2	0.31	0.264	"+bedroom +apt +private +bath +heart"	5525
3	0.253	0.275	"+apartment +beautiful bedroom +studio +private"	3948
4	0.268	0.28	"+cozy +room +bedroom +manhattan +apartment"	6013
5	0.261	0.283	"+spacious +room sunny +apt +bedroom"	6946

Table 2: Text Topics for Name field in the listing

Topic ID	Doc Cutoff	Term Cutoff	Term	Number of Docs
1	0.147	0.062	"+home +guest +stay +day +time"	8335
2	0.17	0.053	"+place +adventurer +solo +solo adventurer +good"	3482
3	0.154	0.059	"+minute +train +walk +station +manhattan"	7770
4	0.13	0.06	"+building +view +gym +rooftop +doorman"	6873
5	0.15	0.061	"+dryer +tv +coffee +microwave +washer"	7575
6	0.153	0.059	"+live room +living +room +bedroom +dine"	9133
7	0.115	0.052	"+distance +walk distance +walking +restaurant +bar"	6975
8	0.133	0.062	"+bar +restaurant +shop +heart +great"	8247
9	0.112	0.059	"+ceiling +high +high ceiling +loft +hardwood"	5931
10	0.114	0.061	"fully +equip +furnish +enjoy +offer"	6949
11	0.116	0.055	"+minute +walk +minute walk +ride +short"	7507
12	0.144	0.061	"+floor +house +entrance +brownstone +garden"	7401
13	0.136	0.06	"+size +bed +queen +full +sleep"	8379
14	0.125	0.059	"bedroom +bedroom apartment +apartment +renovate +building"	7356
15	0.136	0.06	"+private +share +bathroom +private room +room"	8130
16	0.128	0.06	"+central +park +upper +east +subway"	7338
17	0.082	0.055	"â +wireless +internet +tv +flat"	1962
18	0.119	0.06	"+transportation +easy +access +public +easy access"	6413
19	0.088	0.061	"+grocery +grocery store +store +block +laundry"	5271
20	0.133	0.061	"+light +natural +window +lot +natural light"	7411

Table 3: Text Topics of Description Field in listing

Topic ID	Doc Cutoff	Term Cutoff	Term	Number of Docs
1	0.2	0.163	"+smoke +pet +allow +party +pet"	8239
2	0.152	0.172	"+apartment +leave +door +shoe +shoe"	7301
3	0.153	0.158	"+quiet +hour +quiet hour +building respectful"	3422
4	0.174	0.173	"+guest +time +check +allow +check"	7309
5	0.152	0.163	"+loud +music +loud music +noise +neighbor"	3307
6	0.146	0.171	"+home respectful +treat +neighbor +space"	6462
7	0.145	0.161	"+house +shoe +home +shoe +treat"	4536
8	0.141	0.173	"+kitchen +room +clean +bathroom +keep"	5783

Table 4: Text Topics for House Rules field in listing

Topic ID	Doc Cutoff	Term Cutoff	Term	Number of Docs
1	0.422	0.122	"friendly kid family detector laptop friendly workspace"	7971
2	0.297	0.12	"+dish refrigerator silverware stove oven"	8674
3	0.361	0.107	"translation +miss hosting amenity en +stay"	10250
4	0.278	0.11	"intercom wireless intercom wireless buzzer tv"	11268
5	0.296	0.117	"+park free street free street +greet"	9776
6	0.253	0.112	"first kit aid first aid kit card"	10615
7	0.245	0.111	"lock bedroom door bedroom door +allow"	10771
8	0.195	0.109	"+pet live property cat dog"	4718
9	0.177	0.117	"wide access wide doorway doorway clearance"	2483
10	0.182	0.113	"private room +private live room living private entrance"	5981
11	0.209	0.119	"+allow +stay +long term stay long term"	6481
12	0.198	0.119	"elevator doorman gym washer wheelchair"	6798

Table 5: Text Topics for Amenities field in listing

Main Effects:	Estimates:
Number of Bathrooms	28.95
Location Score	33.48
Number of Bedrooms	21.84
Cleaning Fee	0.48
Minimum Nights	-4.06
Accommodation Limit	13.03
Neighborhood (Manhattan)	51.07
Other Property Type vs Bronx	-7.74
Other Property Type vs Brooklyn	1.53
Other Property Type vs Manhattan	25.91
Other Property Type vs Queens	0.48
Entire Home vs Bronx	-30.91
Entire Home vs Brooklyn	30.02
Entire Home vs Manhattan	46.83
Entire Home vs Queens	9.21
Private Room vs Bronx	-11.21
Private Room vs Brooklyn	-4.27
Private Room vs Manhattan	-19.4
Private Room vs Queens	1.77

Table 6: Main Effects and Interaction Effects of significant variables in Price prediction

The MEANS Procedure

Variable	Label	Mean	Std Dev	Minimum	Maximum	N
accommodates	accommodates	2.86	1.87	1.00	16.00	50220
bathrooms	bathrooms	1.14	0.49	0.00	48.00	50118
bedrooms	bedrooms	1.17	0.75	0.00	15.00	50163
beds	beds	1.58	1.09	0.00	40.00	50151
square_feet	square_feet	717.59	596.76	0.00	5000.00	480
security_deposit	security_deposit	290.89	503.32	0.00	5100.00	30693
cleaning_fee	cleaning_fee	65.29	55.46	0.00	975.00	37839
guests_included	guests_included	1.50	1.12	1.00	16.00	50220
extra_people	extra_people	14.17	24.14	0.00	300.00	50220
minimum_nights	minimum_nights	7.67	223.93	1.00	50000.00	50220
maximum_nights	maximum_nights	44232.24	9583606.57	1.00	2147483647.0	50220
number_of_reviews	number_of_reviews	20.52	38.80	0.00	550.00	50220
review_scores_rating	review_scores_rating	93.74	8.44	20.00	100.00	38320
review_scores_accuracy	review_scores_accuracy	9.59	0.84	2.00	10.00	38265
review_scores_cleanliness	review_scores_cleanliness	9.26	1.08	2.00	10.00	38284
review_scores_checkin	review_scores_checkin	9.73	0.72	2.00	10.00	38234
review_scores_communication	review_scores_communication	9.75	0.71	2.00	10.00	38272
review_scores_location	review_scores_location	9.50	0.80	2.00	10.00	38225
review_scores_value	review_scores_value	9.39	0.89	2.00	10.00	38227
calculated_host_listings_count	calculated_host_listings_count	3.67	11.88	1.00	132.00	50220
reviews_per_month	reviews_per_month	1.37	1.62	0.01	19.19	39430
days_since_last_review		225.55	339.48	0.00	3041.00	39430
duration_first_review		22.78	21.84	-31.00	118.00	39424

Table 7: PROC MEANS procedure in SAS Enterprise Guide showing variable value distribution

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_number_of_reviews	Transformed: number_of_reviews	2	1.0000	1.0000	1.0000
LOG_review_scores_cleanliness	Transformed: review_scores_cleanliness	3	0.7549	0.7706	1.0209
LOG_calculated_host_listings_cou	Transformed: calculated_host_listings_count	10	0.4853	0.5241	1.0800
host_response_rate	host_response_rate	3	0.4281	0.4264	0.9962
LOG_review_scores_accuracy	Transformed: review_scores_accuracy	3	0.3114	0.3652	1.1726
LOG_review_scores_value	Transformed: review_scores_value	2	0.2878	0.2912	1.0118
days_since_last_review		2	0.2364	0.1457	0.6162
OVERVIEW_IND		1	0.0781	0.0677	0.8669
REP_require_guest_profile_pictur	Replacement: require_guest_profile_picture	1	0.0729	0.0593	0.8125
jumio		1	0.0652	0.0378	0.5801
LOG_review_scores_communication	Transformed: review_scores_communication	1	0.0650	0.0614	0.9447
work_email		1	0.0647	0.0000	0.0000
TextTopic17_2	_1_0_+apartment,+leave,+door,+shoe,+shoe	1	0.0590	0.0000	0.0000
TextTopic18_4	_1_0_intercom,wireless intercom,wireless,buzzer,tv	1	0.0585	0.0494	0.8442
TextTopic16_10	_1_0_fully,+equip,+furnish,+enjoy,+offer	1	0.0530	0.0000	0.0000

Table 8: Variable importance in the Final Decision Tree

Effect	Variable	Class Level	Standardized Estimate
LOG_NUMBER_OF_REVIEWS	LOG_NUMBER_OF_REVIEWS		0.323131
IMP_HOST_RESPONSE_RATE	IMP_HOST_RESPONSE_RATE		0.12545
NEIGHBOURHOOD_GROUP_CLEANSED_M...	NEIGHBOURHOOD_GROUP_CLEANSED	MANHATTAN	-0.09256
REP_ROOM_TYPE_1	REP_ROOM_TYPE	1	-0.07867
IMP_LOG_REVIEW_SCORES_CHECKIN	IMP_LOG_REVIEW_SCORES_CHECKIN		-0.07484
IMP_LOG_REVIEW_SCORES_CLEANLINES	IMP_LOG_REVIEW_SCORES_CLEANLINES		0.074169
LOG_PRICE	LOG_PRICE		0.06579
IMP_LOG_REVIEW_SCORES_VALUE	IMP_LOG_REVIEW_SCORES_VALUE		0.064905
NEIGHBOURHOOD_GROUP_CLEANSED_B...	NEIGHBOURHOOD_GROUP_CLEANSED	BROOKLYN	-0.04959
GOVT_ID_0	GOVT_ID	0	0.041737

Table 9: Standardized estimates of important variables from LARS model

Odds Ratio Estimates

Effect		Point Estimate
IMP_LOG_review_scores_checkin		0.023
IMP_LOG_review_scores_cleanlines		999.000
IMP_LOG_review_scores_value		999.000
IMP_host_response_rate		10.211
LOG_number_of_reviews		1.950
TextTopic16_12	0 vs 1	0.705
TextTopic18_2	0 vs 1	0.754
TextTopic18_6	0 vs 1	0.739
TextTopic18_9	0 vs 1	0.701

Table 10: Odds Ratio from the Stepwise Regression model