Airbnb, Host to Superhost

Saahil Anil Chawande
schawan@ncsu.edu
North Carolina State University

Roshani Narasimhan
rnarasi2@ncsu.edu
North Carolina State University

Vishva Pradeep Shah
vpshah@ncsu.edu
North Carolina State University

Jeel Sukhadiya
jsukhad@ncsu.edu
North Carolina State University

1 INTRODUCTION AND BACKGROUND

1.1 Problem Statement

Airbnb defines Superhosts as experienced hosts who prove to be an ideal example for other hosts in providing exceptional services and experiences to guests. Using the Airbnb listings data from New York City, our study investigates the relative importance of various criteria that need to be fulfilled in order for an Airbnb host to obtain a Superhost title.

To quantify these analyses, we use the Logit model, Probit model and Lasso regression model to study the importance of the features by evaluating the coefficients produced by these models. Further, an analysis is performed on data after grouping the listings by neighborhood by using the same models, but with additional attributes which are specific and relevant within a single neighborhood.

We seek to summarize the effect and importance of the significant features using bar charts and pie charts. To show the results of the neighborhood-specific study, we will use a bar chart (to compare the contribution of the attributes in making a host.

1.2 Background

In recent years, the development and diffusion of technology have facilitated and enabled the emergence of peer-to-peer online platforms that promote user-generated content, sharing, and collaboration. Since its foundation, Airbnb has grown extremely rapidly and now surpasses the major hotel chains in accommodation offered and market valuation. In the absence of reputational corporate brand identity in Airbnb shared rentals, the hosts must provide some information for purposes of increasing the attractiveness of their property in terms of amenities and other pricing criteria, which are comparable and can be searched.

In the operation of the sharing economy, trust has a pivotal role as consumers do not make their decision based on trusted brand names but in general, they rely on other information about the listings. Travelers can obtain information about accommodations from two main sources. One is the basic information on the website posted by owners and comprising details such as facilities, prices, and photos to help them picture the accommodation; the other source is reviews posted by previous users. The latter seems to be more important since it reflects real experiences and is thus perceived to be more credible.

Airbnb uses a kind of badge system, the “Superhost” badge, designed for accommodation providers.[3] Owners must make continual contributions to the community, such as to inquiry from potential guests quickly, provide good amenities, in order to obtain and keep this badge, which gives them higher status within the Airbnb community of owners.[2]

1.3 Related Work

The growth of Airbnb has attracted substantial academic attention. There is wide consensus in the literatures of tourism, information systems, and digital marketing (e.g., Corbitt, Thanasankit, and Yi 2003; McKnight, Choudhury, and Kacmar 2002; M. J. Kim, Chung, and Lee 2011; Filieri, Alguezaui, and McLeay 2015) that users’ perception of website quality is positively related to their trust in that website. The sharing economy is enabled by community-based technological services to support user contributions to assist online commercial sharing activities. Therefore, the quality of online sharing systems is a crucial technical enabler of such activities.

System quality measures the desired capabilities of an e-commerce system, such as availability, reliability, and ease of use (DeLone and McLean 2003). The importance of system quality in the use of the system is well established in the information systems literature (e.g., Venkatesh et al. 2003; Venkatesh, Thong, and Xu 2012).

In this study, system quality refers to the quality of the service providers(Airbnb hosts) as perceived by the customers of the platform(the guests). Customers feel safer and have greater trust toward a host if they perceive that the latter operates reliably.

The integrative trust formation model developed by McKnight, Choudhury, and Kacmar (2002) contends that institutional assurance can influence individuals’ decisions on information disclosure.

Prior research has emphasized that certifications in the form of trust seals such as VeriSign or TRUSTe can help consumers to trust shopping websites (Hu et al. 2010; D. Kim, Steinfield, and Lai 2008; Xu et al. 2009). In our study, this trust seal refers to the superhost badge awarded by Airbnb to hosts for their exceptional customer service.

These scholars (Roelofsen and Minca 2018, p. 177) have underlined the importance of becoming superhosts in terms of a bigger visibility on the Airbnb platform. Wang and Nicolau (2017) have also discovered a positive relationship between the superhost status and the price, evidencing an increase in the income for the hosts that obtain the badge.

Liang, Schuckert, Law, and Chen (2017), who embed Airbnb’s superhost badge into a gamification framework, come to the same conclusion in showing that Airbnb guests are willing to pay more to a superhost than to a regular host and that a superhost is more likely to receive reviews.

The work by (Gunter, 2017) hypothesizes that the 4 superhost criteria laid by Airbnb are not equally important. He quantifies the marginal contributions of the four superhost criteria, in order to

assess their relative importance. This work uses binary response models such as Logit and Probit in order to test the hypothesis. In his study, he has used data from AirDNA, with a geographic focus on San Fancisco and Bay Area.

Another considered aspect is the localization of Airbnb accommodation in the territories and the impact of Airbnb on cities (Gurran and Phibbs 2017). Dudás et al. (2017) have defined a method to map the spatial distribution of Airbnb accommodations, whereas Gutiérrez et al. (2017) have analyzed their spatial distribution.

Apart from the present study, the studies by Gunter(2017) and Giulia Contu1, Claudio Conversano , Luca Frigau, Francesco Mola (2019) were the only published articles investigating Airbnb superhost criteria in detail. This makes the investigation of the journey from Airbnb host to superhost still a quite novel study in the literature.

While these are a few published articles concentrating on the aspects which constitute to make a host, a superhost; there is currently none exploring the determinants of the superhost status more closely, by geographically localizing the general results, in order to understand which factors are more influential at a granular level, which is what this research addresses.

The present paper builds on the study presented in Gunter (2017) and the objective of this work is to broaden this study to investigate and identify the key determinants from various features that are influential in this promotion of host status by considering additional attributes and models.

The main research question addressed by this work is- what factors influence an Airbnb host in becoming a superhost and how these factors change when we focus on a localized geographical location(ie. a neighborhood).

2 METHOD

2.1 Approach

Out of the several attributes present in the Airbnb New York dataset, we shortlisted the following attributes: Host Total Listings Count, Property Type, Instant Bookable, Neighborhood Group Cleansed, Host Response Rate, Guests Included, Review Scores Rating, Host Identity Verified and Cancellation Policy. Now to determine the weights of each attribute that contribute to the host being a superhost, we used three learning models namely, Logistic (Logit), Probit Regression and Lasso Regression model.

2.1.1 Logistic Regression: Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression). Mathematical, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0(certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name.[4]

This model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, in our case, it is the probability that a host is a superhost or not. The logarithm of the odds is the logit of the probability, the logit is defined as follows:

$$logit(p) = \ln(\frac{p}{1-p}) \quad \text{for } 0 < p < 1 \tag{1}$$

2.1.2 Probit Model: A Probit Model is a type of regression where the dependent variable can take only two values, for example married or not married. This model is a popular specification for a binary response model. As such it treats the same set of problems as does logistic regression using similar techniques. It is most often estimated using the maximum likelihood procedure, such an estimation being called a probit regression.[5]

A compact representation of the Probit model can be shown as follows:

$$Probability(y = 1|X) = \Phi(X\beta) \tag{2}$$

where Φ is a standard normal distribution.

2.1.3 Lasso Regression. Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multi-colinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.[12]

The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator.

Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g. Ridge regression) doesn't result in elimination of coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.

Lasso solutions are quadratic programming problems. The goal of the algorithm is to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

which is the same as minimizing the sum of squares. Some of the βs are shrunk to exactly zero, resulting in a regression that's easier to interpret.

A tuning parameter, λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage:

- When $\lambda = 0$, no parameters are eliminated. The estimate is equal to the one found with linear regression.
- As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when $\lambda = \infty$, all coefficients are eliminated).
- As λ increases, bias increases.
- As λ decreases, variance increases.

2.2 Rationale

The **Logit**, **Probit** and **Lasso** models have been chosen since they allow us to reach the main goal of inferencing the effects of the covariates over the binary response variable Superhost. Furthermore, they evaluate the marginal effects, providing more interpretable results especially for the coefficients of discrete covariates.

We found the logit, probit and lasso regression model to be more appropriate for our study when compared to the following models:

2.2.1 Linear Probability Model. We have often used binary variables as explanatory variables in regressions. What about when we want to use binary variables as the dependent variable which is happening in our case?

It's possible to use OLS:

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \tag{4}$$

where Y is the dummy variable. This is called the linear probability model.

Estimating the equation:

$$\widehat{P}(y = 1|x) = \widehat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + ... + \widehat{\beta_k} x_k \tag{5}$$

where $\widehat{y}$ is the predicted probability of having y = 1 for the given values of $x_1, x_2, ..., x_k$.

Problems with the linear probability model (LPM):

- Heteroskedasticity: can be fixed by using the "robust" option which is not a big deal.
- It is possible to get $\widehat{y} < 0$ or $\widehat{y} > 1$. This makes no sense since we can't have a probability below 0 or above 1. This is a fundamental problem with the LPM that we can't patch up.

Solution: Use the logit or probit model. These models are specifically made for binary dependent variables and always result in $0 < \widehat{y} < 1$. The following graph shows where the LPM goes wrong and the logit / probit model works:

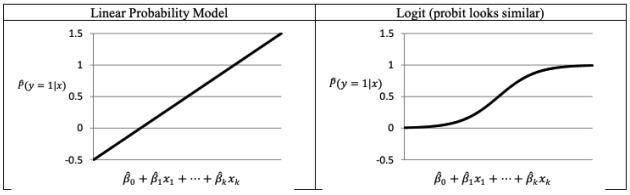

Figure 1: Comparison between LPM and Logit / Probit

This is the main feature of a logit/probit that distinguishes it from the LPM - predicted probability of y = 1 is never below 0 or above 1, and the shape is always like the one on the right rather than a straight line.

2.2.2 Comparison between Lasso and Ridge Regression. In the case of Machine Learning, both ridge regression and lasso regression find their respective advantages. Ridge regression does not completely eliminate (bring to zero) the coefficients in the model whereas lasso does this along with automatic variable selection for the model. This is where it gains the upper hand. While this is preferable, it should be noted that the assumptions considered in linear regression might differ sometimes.[13]

Both these techniques tackle over fitting, which is generally present in a realistic statistical model. It all depends on the computing power and data available to perform these techniques on a statistical software. Ridge regression is faster compared to lasso but then again Lasso has the advantage of completely reducing unnecessary parameters in the model.

In our case for the Airbnb dataset, we want to consider only the top features that would make more contribution to the results, hence Lasso regression helps us achieve that in a more accurate manner.

3 EXPERIMENTS

3.1 Dataset

For the proposed approach we use the New York City Airbnb data, provided by Airbnb at the insideairbnb.com website. The dataset consists each and every possible details of the listings mentioned on the Airbnb website. The dataset contains 106 types of attributes, with 50,599 instances. Some of the attributes closely resemble each other and are categorical or a super set representative of each other. We intend to choose a tightly packed set of attributes which could meet the requirements of the goal. We choose a set of 9 attributes which could potentially attribute to the conversion of a host to superhost. The correlation matrix between the 9 attributes of the city wide data is displayed in the Fig[3]. With the goal of studying trends on a more granular level, we used the proposed methods to analyse trends on neighborhood wise data of New York City. We increased the number of attributes to 14, by adding a couple of attributes more closely associated to individual neighborhoods, to get detailed results about neighborhood wise trends. The number of neighborhood wise-listings is listed in the Table[1]. The number of superhost for each neighborhood is accurately described in the Fig[2]

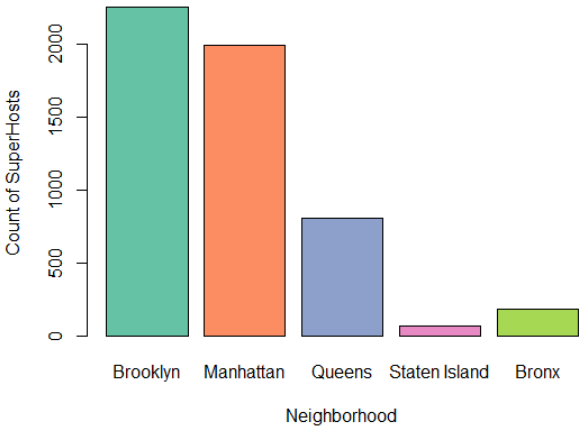

Figure 2: SuperHost division by neighborhood

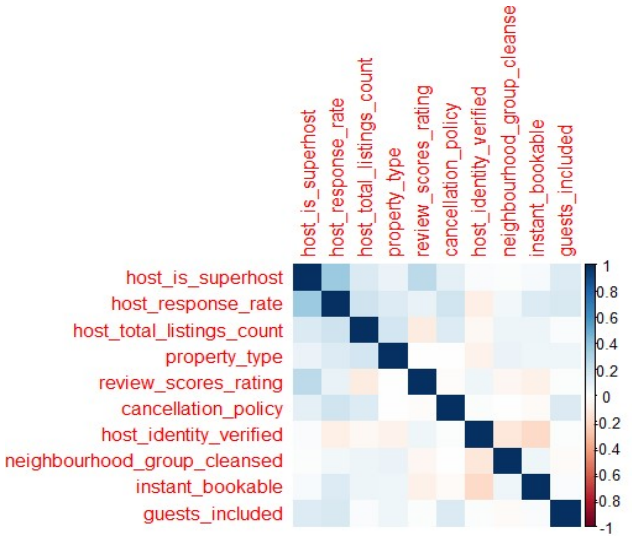

Figure 3: Attribute correlation plot

Neighborhoods	No of Listings
Manhattan	21724
Brooklyn	20436
Queens	6093
Bronx	1195
Staten Island	368

Table 1: Distribution of Airbnb listings across Neighborhoods

3.2 Hypotheses

The primary goals of our exploratory analyses are to identify the attributes which influence the status of hosts and to evaluate the relative order of importance of such attributes.

It is reasonable to believe that prospective guests deliberating over a particular property, tend to provide consideration to the reviews and ratings provided by previous guests who have booked that property(Gunter, 2017). Additionally, hosts who actively respond to prospective guests' queries and reveal information regarding the listings in question, might attract more guests compared to others. Thus the involvement and responsiveness of the host might be re-assuring to the prospective guests. We wish to explore whether the responsiveness of a host and the reviews of their listings help them garner the 'superhost' status.

Guests might also find it important that they have some flexibility in terms of cancelling their booking, even closer to their trip. Thus intuitively, we expect that flexibility in cancellation policy followed by a host should contribute significantly to whether he/she becomes a super host. Thus, we would be analyzing the impact of cancellation policy to the hosts' status.

We explore other factors which could lead to a host becoming a super host - such as the reliability of the host. We hypothesize that hosts who have their identity verified by Airbnb tend to make the guests assured of their legitimacy and the guests feel more welcome to book their listings.

We expect that neighborhood or property type might not play a major role in hosts becoming super hosts. It seems more logical to consider that the property should satisfy the requirements of the prospective guests and also have a good standing in terms of rating.

On similar lines, it seems reasonable to expect that the price of the properties are matched according to the market demand and amenities offered by the properties. If so, two listings - one provided by host and another by a superhost, having comparable services and within the same neighborhood will most likely have similar price ranges. Previous research (Gunter, (2017)) suggests that Airbnb listings offer competitive (most often cheaper) prices as compared to hotels, with almost same level of services and amenities. Thus it would be logical that the pricing gets balanced among Airbnb listings based on locations, type of property and other services provided. We wish to explore whether pricing of listings provided by hosts and superhost differ and if they do, are they significant enough to contribute to a change of status. However, as pricing is dependent on individual neighborhoods and the ease of access of amenities such as hospitals, schools and transport within the neighborhood, it makes more sense to evaluate price in the context of one neighborhood. (Guttentag, D. (2015))

This leads us to another important novel research aspect - whether and how the importance of attributes such as price of listings, number of bedrooms, type of rooms, etc vary across the locations within New York City, namely - Bronx, Brooklyn, Manhattan, Queens and State Island. We wish to explore these trends for each neighborhood and estimate how they affect the status of a host within that particular neighborhood.

We hypothesize that some attributes such as number of bedrooms or type of property or the level of privacy offered by the property, could assume more importance than the global criteria across NYC, such as the reputation of the host, thereby altering the expected trends.

This analysis would be interesting to lead us to a more specific reasoning of whether hosts in a particular neighborhood can focus on some particular aspect of their listing, so as to have higher probability of achieving the 'superhost' status.

3.3 Experimental Design

3.3.1 Steps involved in pre-processing. The first step of pre-processing is to handle the rows which do not have a meaningful class label of whether a host is superhost. We decided to remove rows with missing values of class labels (essentially, we removed rows with blanks and N/As). After this step, we are left with 50036 rows.

We performed the following transformations on the attributes to make them better suited for the analyses:

- Cancellation policy in the data had six classes, namely - 'flexible', 'moderate', 'strict_14_with_grace_period', 'super_strict_30', 'strict', 'super_strict_60'. For the purpose of this study, we consider flexible and moderate as 'lenient' and represent them with 0 or not "strict". We consider all the other classes as "strict" and relabelled them as 1.

- We assigned numeric class values to the categorical attribute neighborhood, i.e relabelled 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island' and 'Bronx' as 0,1,2,3 and 4 respectively.
- Similarly, for property type, we have replaced the original labels with numeric values to the 35 different types of properties.
- For blanks and NAs in the host response rate attribute, we replaced them with the average of this attribute across positive and negative classes. When the host is a superhost, the average response rate is 89% and when the host is not a superhost, the average in 79%. We did not replace all the values with one average value, on account of the imbalance in the positive and negative classes.
- We converted the attributes host_is_superhost, instant_bookable and host_identity_verified from logical type to numerical type and host_response_rate from character to numeric type (i.e, 0 or 1).
- We replaced the review scores rating and the total listings count with the averages in the same way as the host response rate.
- We normalized the attribute host_total_listings_count to bring all the values to the range of 0 to 100. (Originally, this attribute had a range of 0 to 1767, for our dataset.)

3.3.2 Model Training. We divided the data into training and testing data in the ratio of 60:40. We used the training set with 9 attributes and the class label to train the Logit, Probit and Lasso regression models. We obtained the relative importance of the three models based on the coefficient of the models. We compared the predictions from the three models with the actual values of the class label to determine the accuracy of the models.

For the Lasso model, we check a range of Λ (weight of the penalty term) values from 10^{-3} to 10^{2} and choose the $\Lambda = 0.001$ which minimizes the error. We train the model again on this Λ value.

3.3.3 Software tools and major packages used. All the above mentioned steps have been performed in R, using the following packages - glmnet (for glm implementation of Logit, Probit and Lasso models), dplyr (for select), anchors (for replace.value) , stringr (for string manipulations) and caTools (for sample.split).

3.3.4 Inclusion of attributes and their pre-processing for neighborhood specific analysis. As mentioned before, we included five more attributes, which make more sense when considered within a neighborhood, in order to understand the trends within neighborhoods. These attributes are - room_type (Entire apartment, shared room, private room or hotel room), number of bathrooms, number of bedrooms, number of beds, price of listing and charge per extra guest. We pre-processed the extra attributes as follows:

- We converted room_type from categorical to numeric, i.e. 0,1,2,3 to facilitate our analysis.
- We removed rows which do not have values for number of bathrooms, number of bedrooms, number of beds (averaging over the rest of the data items to replace N/A values for these attributes did not seem like the best idea). Post this step, we are left with 49816 rows, which still makes a substantial number of data items.
- We normalized price and extra charge per extra person (which are originally listed in dollar amounts) to be in the range of 0 to 100, to bring some uniformity in the data.

Then the main dataset is divided into 5 smaller datasets, each corresponding to one of the five neighborhoods of 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island' and 'Bronx'.

3.3.5 Model Training. The training process followed for individual neighborhoods was similar what was followed for the city-wide analysis. We divided the data into training and testing data in the ratio of 60:40. We used the training set with 14 attributes, after excluding the neighborhood attribute and the class label- host_is_superhost to train the Logit, Probit and Lasso regression models. We compared the predictions from the three models with the actual values of the class label to determine the accuracy of the models for each neighborhood. We obtained the relative importance of the three models based on the coefficient of the models, for each neighborhood.

4 RESULTS AND DISCUSSION

4.1 Results

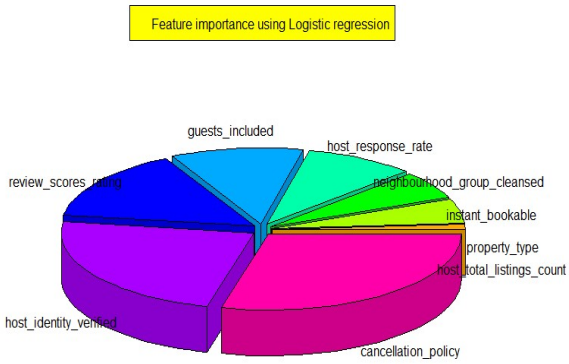

Figure 4: Feature Importance Using Logistic Regression

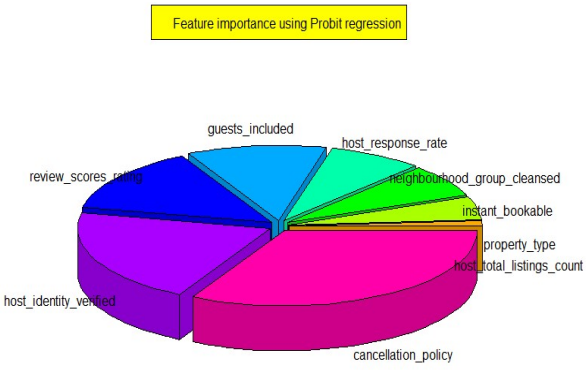

Figure 5: Feature Importance Using Probit Regression

4.1.1 New York City Data Analysis. The proposed methods were used to find out the important factors which could be responsible

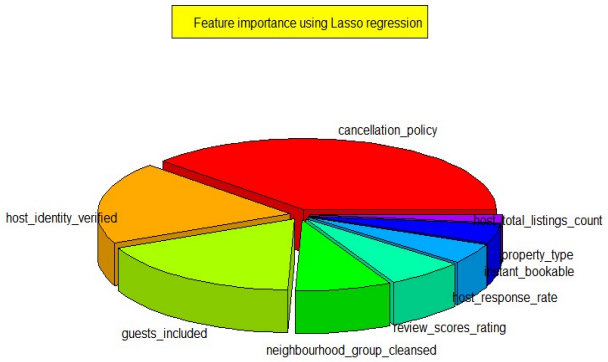

Figure 6: Feature Importance Using Lasso Regression

Features	Logit	Probit	Lasso
cancellation_policy	0.3549	0.227	0.05465
host_identity_verified	0.28686	0.14128	0.0269
review_scores_rating	0.18527	0.0972	0.00936
guests_included	0.1365	0.0809	0.02551
host_response_rate	0.136	0.05477	0.00684
neighborhood_group_cleansed	0.07322	0.0485	0.01164
instant_bookable	0.06351	0.03414	0.00629
property_type	0.01283	0.00742	0.00266
host_total_listings_count	0.000328	-0.000239	0.0

Table 2: Feature Importance using Logit, Probit and Lasso models

for conversion of an Airbnb host to superhost. These methods were applied on the New York city data as well as on the neighborhood specific data of New York city to analyse localized trends. The study performs the analysis on 9 potential city wide attributes while the study exploring the neighborhood trends performs the analysis on a total of 14 different kinds of potential attributes. The accuracy values of the three methods: logistic regression, probit regression, lasso regression are shown in Table[1]. All the three models deliver similar results, marginally different from one another. The aim of our study is to study the factors responsible for the conversion of host to superhost. These results are only applicable to the dataset that has been chosen to perform the study, the results may vary from dataset to dataset. The relative importance of these attributes, obtained after applying these models on the entire New York City data has been represented in the Fig[4], Fig[5] and Fig[6]. Based on the results, we find out that Cancellation policy is one of the most important factors for a host to become a superhost.

4.1.2 Neighborhood-wise analysis of New York City. The study provides more detailed analysis of the factors that might be responsible for the superhost status by analysing localized trends in particular neighborhoods. The study takes into account the five major

Methods	Accuracy(%)
Logistic Regression	79.163%
Probit Regression	79.586%
Lasso Regression	79.924%

Table 3: Accuracy comparison Chart

neighborhoods of New York city, which is Bronx, Brooklyn, Staten Island, Manhattan, Queens. Similarly to the methods applied to the dataset corresponding to New York City, we apply the three binary response models to analyse localized trends in the data corresponding to these neighborhoods. The Fig[7], Fig[8] and Fig[9] describe the trends we found out on the neighborhood level. The study took into account a total of 14 different attributes for each method used, but for simplicity, we display only the top four attributes per neighborhood, which are the most responsible for the superhost status. The results obtained by using the logistic regression method

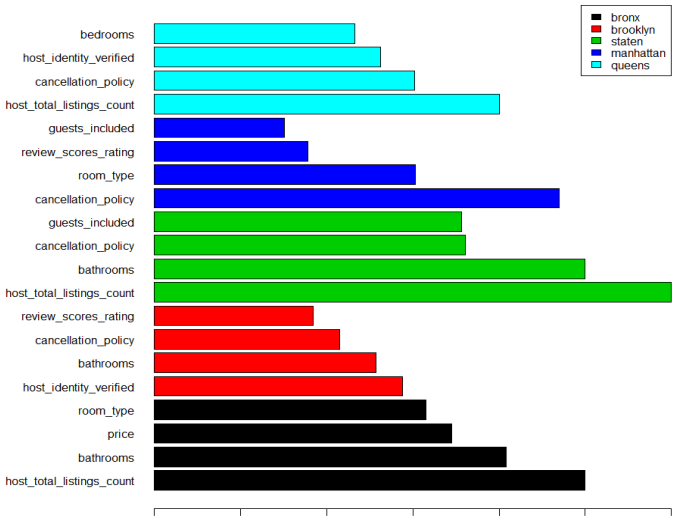

Figure 7: Neighborhood wise feature importance Using Logistic Regression

on the neighborhood data are displayed in Fig[7], based on the distribution of Airbnb listings across different neighborhoods indicated in Table[1]. It is evident that in neighborhoods with more number of listings like Manhattan, Brooklyn and Queens common attributes such as cancellation_policy, review_scores_rating and host_identity_verified contribute to the Airbnb Superhost status. The results thus align with the results obtained on the entire city data, as most of the observations are from these neighborhoods. The neighborhoods of Bronx and Staten Island have relatively less number of listings and hence the resultant important attributes are more related to the properties specific to the listings like the total listings provided by the hosts, number of bathrooms, price etc.
The results represented in the Fig[7], Fig[8] and Fig[9] only show a means of attribute dominance in one particular neighborhood, results from one neighborhood do not correlate with another, they are independent of each other. The relative ordering of top factors

contributing to the conversion of the host to superhost is similar for the Probit model and Lasso models, wherein, guests in bigger neighborhoods focus on service and hospitality and guests smaller neighborhoods focus on the attributes particular to the listing.

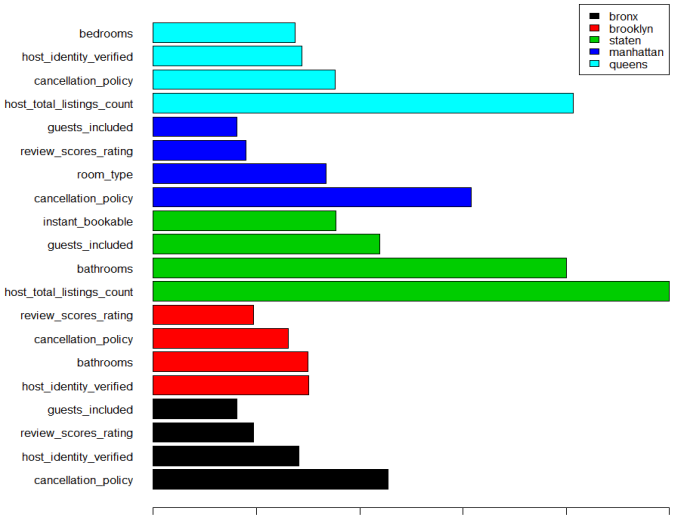

Figure 8: Neighborhood wise feature importance Using Probit Regression

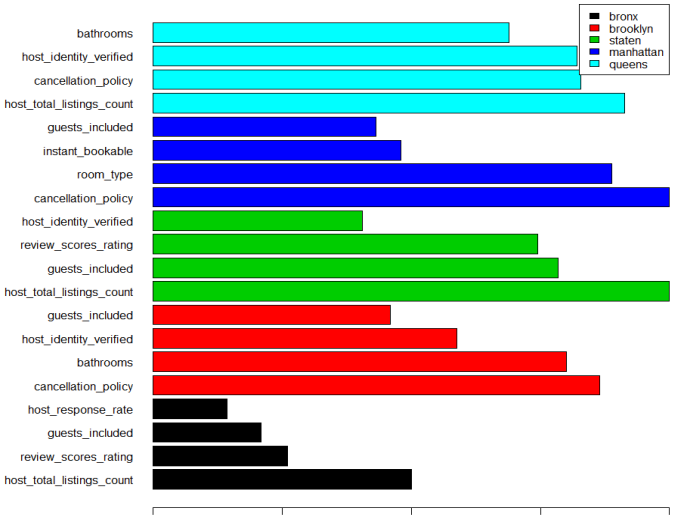

Figure 9: Neighborhood wise feature importance Using Lasso Regression

4.2 Discussion

4.2.1 General trends across New York City. It is evident from the results mentioned above, that the rating of the listing, based on the reviews from previous guests, plays an important role in the current standing of the hosts. It gives a sense of guarantee to the prospective guests that the amenities and services offered by the host are above par.

On the same lines, the whether or not the identity of the host is verified by Airbnb seems to be significant, since this confirms the legitimacy of the host. The host's responsiveness also carries some weight as prospective guests who are looking for an accommodation would prefer answers to their queries, to help them in making a decision regarding which accommodation to choose. As hypothesized, it seems like the involvement and responsiveness of the host gives a guarantee that the transaction with the host will remain a smooth experience will be ensue.

Cancellation policy has a very a significant contribution to the experience of a guest, as seen from the results. As opposed to what we expected, guests seem to prefer hosts with stringent cancellation policy. Guests seem to interpret the strict cancellation policy as reassurance that the host will also not cancel the booking from their end. A simple analysis of the cancellation policy within the dataset supports this result - about 58% of superhosts have strict cancellation policy compared to only 43% of hosts having strict cancellation policy. Previous research by (U. Gunter, 2017). also validates this significance of stringent cancellation policy in achieving the superhost status.

Based on our results, attributes which are more associated to the listings such as property type, whether or not the listing is instantly bookable and number of listings of the host puts out - do not account much to the superhost status. These attributes do not significantly qualify the attributes of the host such as their reliability or promptness or the level of services offered by them, thus do not have much importance in making them a superhost.

4.2.2 Trends across the neighborhoods. From the weights of the attributes, given by the models, we realized that across Brooklyn the priority of the guests aligns with the trend of the city. The top five positively impacting factors remain the same, with minor changes to their relative order. Rating, cancellation policy and verification of the host's identity are factors what guests across Brooklyn seek the most. However, the number of bathrooms offered in the listing seems to have a significant negative impact towards the conversion of the host to superhost.

Similarly, Airbnb guests across Manhattan seem to place a lot of importance on the privacy offered by the listing (as denoted by room_type) as well as the number of guests that can be accommodated in the listing. This is in addition to the importance placed on the cancellation policy for the listing and their review scores.

Across Queens, number of bedrooms and number of listings offered by the host negatively impacts the host status, thereby suggesting that guests are looking for listings with lesser number of bedrooms. These guests also place importance on the cancellation policy and the authenticity of the host.

A surprising result across Staten Island, is that guests here place a lot of importance of the number of bathrooms offered in the listing. Since there are very few hosts across Staten Island, one host offering many listings becomes a monopoly and thus tends to become a superhost.

For hosts across Bronx , the level of privacy offered and the price seem to negatively affect the status, the underlying reason could be that the privacy offered, along with the services, might

not be worth the price. Guests preferring Bronx place importance of the number of bathrooms as well. Since there are very few hosts across Bronx as well, one host offering many listings becomes a monopoly and thus tends to become a superhost.

We notice that the price of the listing does not play a major role in most neighborhood, since we estimate that it will be regulated by the market demand and will be specific to the type of property and the neighborhood. Thus, as expected, price of the property does not play an important role. Previous research also indicates price as a superfluous factor in making a superhost.

5 CONCLUSION

Drawing on a large cross-sectional data set of more than 50,000 Airbnb listings from New York City picked from insideairbnb.com, the present study quantifies the contributions of the various attributes to make a host, a superhost and analyzes their relative importance. The findings of our study provide meaningful business insights and policy implications for Airbnb hosts to expand successfully.

The main findings of this study were that in New York City overall, the attributes which contribute the most in promoting a host are cancellation policy, whether the host identity is verified, the host's ratings and the number of guests permitted on the listing site. Thus, Airbnb hosts who want to obtain(or maintain) the superhost status are therefore advised to enforce cancellation policies which ensure guest bookings will not be altered from host-end, focus relatively more of their energy on getting (and keeping) excellent ratings while they must not neglect the remaining criteria.

The results were consistent across the models specified, while achieving over 79% accuracy in each model .

Further, the geographically localized study has also revealed that although in principle, nearby destinations might appear very similar, there are many differences to be highlighted, to further motivate more hosts to make smart business decisions, based on analysis specific to their neighborhood.

We found that in the neighborhoods- Manhattan, Brooklyn and Queens, where the number of Airbnb listings is very high and customers have ample options suiting their preferences, the general attributes of hospitality - such as host response rate, overall reviews of host, flexibility offered, etc prove to be most important; even over the listings site's details itself.

However, in small neighborhoods such as Staten Island and Bronx, where the number of listings is low, the attributes which mattered the most were related to the properties and aesthetics of the listing itself – such as number of bedrooms, number of bathrooms etc.

To sum up, to become a superhost it is necessary to offer a high quality service, support guests right from the booking stage and provide an experience more than the average that is offered by other hosts in the neighborhood. The results of this study have some limitations, due to the small timeline of this project, we were only able to localize results for New York City and compare them to the global results. Future work can be addressed towards a broader geographic area and it would also be interesting to analyze more variables to understand how the results are different for commercial Airbnb hosts versus regular Airbnb hosts.

6 REFERENCES

[1] Ulrich Gunter. 2017. What makes an Airbnb host a superhost? Empirical evidence from San Francisco and the Bay Area. Tourism Management 66 (2017), 26–37.
[2] Giulia Contu, Claudio Conversano, Luca Frigau, and Francesco Mola. 2019. Identifying factors affecting the status of superhost: evidence from Sardinia and Sicily. Quality & Quantity (August 2019).
[3] Sai Liang, Markus Schuckert, Rob Law, and Chih-Chien Chen. 2017. Be a Superhost: The importance of badge systems for peer-to-peer rental accommodations. Tourism Management 60 (2017), 454–465.
[4] 2020. Logistic regression. (March 2020). Retrieved April 3, 2020 from https://en.wikipedia.org/wiki/Logisticregression
[5] 2020. Probit model. (March 2020). Retrieved April 3, 2020 from https://en.wikipedia.org/wiki/Probit_model
[6] Corbitt, Brian & Thanasankit, Theerasak & Yi, Han. (2003). Trust and E-Commerce: A Study of Consumer Perceptions. Electronic Commerce Research and Applications. 2. 203-215. 10.1016/S1567-4223(03)00024-3.
[7] Delone, William & McLean, Ephraim. (2003). The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. J. of Management Information Systems. 19. 9-30. 10.1080/07421222.2003.11045748.
[8] Mcknight, D. & Choudhury, Vivek & Kacmar, Charles ("Chuck"). (2002). The Impact of Initial Consumer Trust on Intentions to Transact with a Web Site: A Trust Building Model. The Journal of Strategic Information Systems. 11. 297-323. 10.1016/S0963-8687(02)00020-3.
[9] Roelofsen, Maartje. (2018). Performing "home" in the sharing economies of tourism: the Airbnb experience in Sofia, Bulgaria. Fennia. 196. 10.11143/fennia.66259.
[10] Wang, Dan & Nicolau, Juan. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. International Journal of Hospitality Management. 62. 120-131. 10.1016/j.ijhm.2016.12.007.
[11] Gurran, Nicole & Phibbs, Peter. (2017). When Tourists Move In: How Should Urban Planners Respond to Airbnb?. Journal of the American Planning Association. 83. 80-92. 10.1080/01944363.2016.1249011.
[12] Stephane. (2015). Lasso Regression: Simple Definition. https://www.statisticshowto.com/lasso-regression/
[13] Abhishek Sharma. (2018). Ridge Regression vs Lasso: How these 2 popular ML Regularization Techniques Work. https://analyticsindiamag.com/ridge-regression-vs-lasso-how-these-2-popular-ml-regularisation-techniques-work/
[14] Guttentag, D. (2015). Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. Current Issues in Tourism, 18, 1192e1217.
[15]http://insideairbnb.com/get-the-data.html

7 ACKNOWLEDGMENTS

We Acknowledge Dr Thomas Price and the Teaching Assistants Mr. Yang Shi and Ms. Ge Gao for their guidance in the project. hyperref

8 OUR WORK

Github link

9 VIRTUAL MEETINGS

Virtual Meetings schedules	Attendance
8th Apr 2020, 2pm-6pm	Jeel,Vishva,Saahil,Roshani
10th Apr 2020, 10am-2pm	Jeel,Vishva,Saahil,Roshani
15th Apr 2020, 3pm-6pm	Jeel,Vishva,Saahil,Roshani
20th Apr 2020, 1pm-5pm	Jeel,Vishva,Saahil,Roshani