Final Project

Noah Costa and Ezra Odio

Introduction

All across the world, the hospitality industry is essential to many people's livelihoods. Many people first think of hotels when hospitality is brought up, but there is another, fast-growing accommodation service that is making waves: Airbnb. As more and more people look to invest in creating Airbnb properties for financial gain, it is essential that these property owners understand what factors make their guests happy in order to ensure success. This report seeks to come up with a model to predict success in terms of guest satisfaction based on a number of different factors related to the Airbnb property.

Data and Data Cleaning

The data is taken from a dataset called "Airbnb Cleaned Europe Dataset" on Kaggle. This dataset is a cleaned version of a dataset used by Kristóf Gyódi and Łukasz Nawaro in their research paper titled "Determinants of Airbnb prices in European cities: A spatial econometrics approach." The original data was collected by executing search queries on Airbnbs platform for certain major European cities.

As the dataset on Kaggle had already been cleaned, not much was necessary in terms of data cleaning, as there were no observations which needed to be removed. The one thing we did do was create a new variable called "satisfied" that was set to 1 if guest satisfaction was greater than 90, and 0 otherwise. This threshold of 90 can be changed depending on the preferences/goals of the potential owner.

Data Citations

Kaggle: https://www.kaggle.com/datasets/dipeshkhemani/airbnb-cleaned-europe-dataset

Kristóf Gyódi, Łukasz Nawaro, Determinants of Airbnb prices in European cities: A spatial econometrics approach, Tourism Management, Volume 86, 2021, 104319, ISSN 0261-5177, https://doi.org/10.1016/j.tourman.2021.104319.

Relevant Variables:

Guest Satisfaction: Overall rating of the listing on scale of 20-100

City: City where Airbnb is located

Price: Full price of accommodation for two people for two nights in Euros

Room.Type: The type of the accommodation. Including shared, private or house/apartment

Superhost: Whether the host is a superhost or not. Superhost definition can be found on Airbnbs website

Multiple.Rooms: Whether the host has 2-4 listings

Business: Whether the host has more than 4 listings

Cleanliness.Rating: Overall cleanliness rating on scale of 2-10

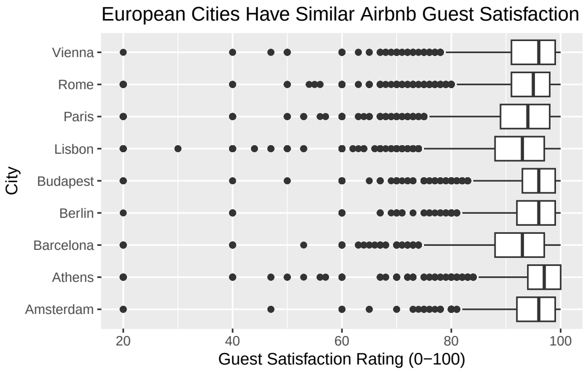
Bedrooms: Number of bedrooms

City.Center..km.: Distance from city center in km

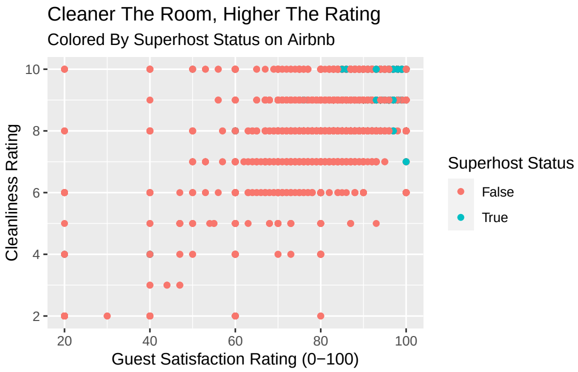
Metro.Distance..km.: Distance from nearest metro station in km

Attraction.Index: Attraction index of the listing (farther from attractions, lower the value)

Exploratory Data Analysis



Looking at the graph above, we can see that the median Airbnb guest satisfaction ratings for the cities in the dataset are all very similar to each other. Additionally, the median Airbnb ratings are all greater than 90.



Looking at this graph, we can see that higher cleanliness ratings seem to correlate with higher guest satisfaction ratings. Additionally, we can also see that Airbnb listings posted by superhosts tend to have higher cleanliness ratings and guest satisfaction ratings.

Methodology

The first step we took to create our models was variable selection. The dataset originally had 19 columns, however 2 were an expanded version of another column (private room and shared room were included in room type) and 2 more were simply the normalized versions of 2 other columns we had (Attraction Index and Restaurant Index). This leaves us with 14 predictor variables and one variable needed to predict. As the introduction mentioned, the variable we will be predicting is Guest Satisfaction. In order to determine which variables would be a part of our model, we used all subset selection on these 14 variables. The City variable has 9 subcategories, so it was broken into 8 dummy variables with Amsterdam being the reference variable, and the Room.Type variable has 3 subcategories so it was broken into private room and shared room dummy variables with home/apartment as the reference variable, and cleanliness rating was treated as a categorical variable rather than continuous so that had 7 additional dummy variables with a score of 2 as the reference. This means there was a maximum of 29 variables that could be in the model. After running all subset selection and checking against with adjusted R^2 and Mallow Cp values, both returned that we should use a model with 21 of these 29 variables. By using the which function we found that 5 of the 8 city dummy variables were to be included as well as 5 out of 8 cleanliness scores, so we decided to include the City variable and cleanliness rating in our final model. The rest of the variables included in the model were Price, Room.Type, Superhost, Multiple.Rooms, Business, Bedrooms, City.Center..km. and Attraction.Index.

As for choosing a model to use these variables in, we wanted to be able to account for the independence concerns between cities while also being able to accurately predict whether or not a guest is satisfied. The best way we thought of to tackle the issue of independence was by using a mixed model for the different cities that our data is taken from. In this type of model, City would be treated as a grouping variable which helps us account for the variability between the different cities, in this case the 9 that are in the dataset. This allows the model to "borrow information" about the slopes between the cities. This is important because our independence assumption is likely violated between cities as users would likely attach how they enjoyed their trip to the overall satisfaction with the Airbnb which is heavily connected to the city itself. However, because we are creating this model for Airbnb owners who want to improve their satisfaction, we want to have high predictability which we cannot determine for a mixed model using our current statistical knowledge.

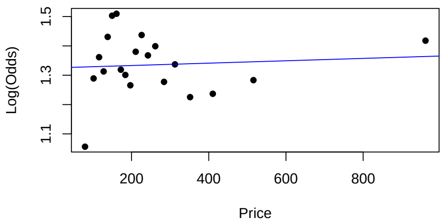
This led us to create a logistic regression model with the data, as we could easily determine if it had a strong predictability. For this model we used the same predictor variables, although now we added in 3 additional interaction terms in order to account for any effect of one variable on another. The three interaction terms we added in were between City * Price, City * City Center Distance, and City * Attraction Index. The reason we chose these interaction terms is because we felt that the relationship between satisfaction and price, city center distance, and attraction index might depend on which city the Airbnb is in. For example, prices may increase in nicer and more popular cities, being closer to the city center might be more important in cities with worse public transit and walkability, and the relationship between satisfaction and attraction index might change depending on how many tourist attractions the city has.

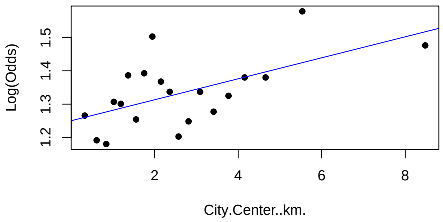
Now that we have all of our variables used to predict guest satisfaction, we need to turn the Guest.Satisfaction variable into a binary outcome. To do this, we created a new variable that we set to 1 if Guest.Satisfaction was greater than or equal to 90, and set to 0 otherwise.

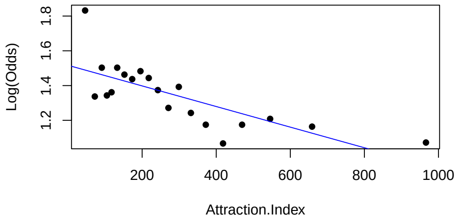
Assumptions

The independence assumption for our logistic regression model does not seem to be satisfied, as we might expect that Airbnbs in different cities share similar qualities. However, given that a lot of these cities seem to share similar qualities, we decided to proceed with our analysis while keeping this possible independence violation in mind.

To test the linearity assumption of our logistic regression model, we created the empirical logit plots below. The empirical logit plots demonstrate that there is a clear linear relationship between the log-odds of a guest being satisfied and the continuous predictors of price, distance to city center, and attraction index.







Results

The final logistic regression model can be seen below:

satisfied ~ City + Price + Room.Type + Superhost + Multiple.Rooms + Business + as.factor(Cleanliness.Rating) + Bedrooms + City.Center..km. + Attraction.Index + City * Price + City * City.Center..km. + City * Attraction.Index

Below is a confusion matrix displaying our model's effectiveness at predicting whether guests were satisfied or not.

```
                  0     1
Not satisfied  4339  1756
Satisfied      4363 31256
```

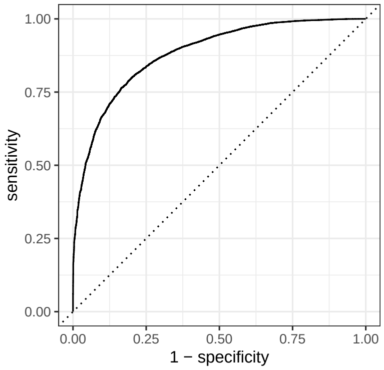
The specificity is 4339/(4339 + 4363) = 0.499.

The sensitivity is 31256/(31256 + 1756) = 0.947.

The positive predicted value is 31256/(31256 + 4363) = 0.878.

The negative predicted value is 4339/(4339 + 1756) = 0.712.

The associated ROC curve for this model can be seen below. This ROC curve has an area under curve of 0.882.



Analysis

The model's high sensitivity rate indicates that our model is good at predicting Airbnb properties with high guest satisfaction as such. The positive predicted value of 0.878 means that we are correct 87.8% of the time when predicting high guest satisfaction. Combined with the ROC curve with AUC = 0.882, we are confident in our model's ability to predict successful Airbnb listings. This achieves our main goal for this project: predicting whether Airbnb listings will be successful in terms of guest satisfaction. If a prospective Airbnb owner wished to predict the guest satisfaction at a potential listing, he/she could feel confident using our model to predict this.

On the other hand, the specificity and negative predicted value are both relatively low. This indicates that the model is not great at predicting which Airbnbs will not satisfy guests. Since we are trying to predict which properties will make guests satisfied, though, this downside of our model is not as relevant.

In terms of which statistically significant (the ones we are confident are not 0 at a significance level of 0.05) slope coefficients seem to be having the biggest effect in terms of predicting guest satisfaction, there are a few that stand out: Price, SuperhostTrue, Room.TypeShared room, Multiple.Rooms, Business.

Price has a coefficient of 1.344e-3, meaning that the odds of a guest being satisfied is predicted to be multiplied by 1.001 with each 1 Euro increase in price of a 2 night stay for 2 people with everything else held constant. While this seems small, it the odds of a guest being satisfied is predicted to be multiplied by 1.477 with a 300 Euro increase in this price.

Similarly, the odds of a guest being satisfied is predicted to be multiplied by 8.14 if the host is a superhost compared to not being one, 0.61 if the host has 2-4 listings compared to one listing, 0.344 if the host has 4+ listings compared to one listing, and 1.92 if the room type is a shared room compared to "Entire home/apt" with everything else held constant.

Discussion

What We Learned

We were pleased to study the results of our model as it seems to be a strong predictor of whether or not a guest will be satisfied in a given Airbnb based on different characteristics of the listing. When trying to determine the key factors to guest satisfaction as an Airbnb host, our model displays clear and significant results for hosts to focus on when listing their next room. The variable that predicted the highest odds of a guest being satisfied was whether or not the owner of the Airbnb was a Superhost. When a host becomes a superhost our model predicts that a guest is 8.14 times more likely to be satisfied with their stay, with all other variables held constant. This is a significant leap in satisfaction rating to where it should strongly encourage any host to strive to become a superhost. The model shows that it would result in a great benefit to the guest and subsequently the host themselves.

Another interesting observation was that the model predicted more expensive Airbnbs to have higher odds of guest satisfaction than less expensive Airbnbs with everything held constant. This observation seems to be a little counterintuitive, and is something that should be explored further. One possible explanation is a psychological one: people might be more predisposed to being satisfied with things that they pay more for.

Next, the model predicts that Airbnbs listed by hosts with 2-4 listings and hosts with 4+ listings have lower satisfaction compared to Airbnbs listed by hosts with only 1 listing with everything else held constant. Perhaps this is due to hosts with multiple listings not being able to devote their full attention towards each group of guests.

Lastly, the model predicts that Airbnbs with shared rooms are almost twice as likely to achieve guest satisfaction as Airbnbs that are the entire home or apartment with everything else held constant. Much like price, this seems to be a little counterintuitive, but a possible explanation is that people generally enjoy being around others.

These findings indicate that a potential host interested in achieving guest satisfaction may want to emulate the behavior of a superhost rather than businesses, list Airbnbs with shared rooms rather than entire homes or apartments, and increase the price of their listings.

Limitations and Improvements

One limitation is that there are many possible variables that we would be interested in that are not in the dataset. For example, the crime statistics in the neighborhood the Airbnb is in, or the year the property was built are both things that could have huge impacts on guest satisfaction.

One way that we could improve our model would be to get a dataset of Airbnb listings from another country or city in Europe and test the effectiveness of our model on it. This would allow us to better understand how generalizable and effective our model is.

Issues of Validity

When considering possible issues pertaining to the model, there is the possibility that the independence assumption is violated between cities. This would be because the 9 cities in the dataset all have the same attractions, food, and cleanliness which may influence the satisfaction of guests as the rating may be a reflection of the city as well as the Airbnb. However, when looking back at our ROC curve we found that it had an area under the curve of 0.882 which implies that our model is a very strong fit of the data. As well as finding a positive predicted value of 87.8% shows that this model is very accurate at predicting what characteristics will lead to a guest being satisfied. One downside to the model is that it is not very accurate at predicting when a guest will not be satisfied, however, that is not the purpose of our research. If a host were to follow our advice, they would be in a much better position to leave a guest satisfied. Another potential issue with the data is that pertaining to the shared rooms. As previously mentioned, renting out a shared room will almost double the odds of a guest feeling satisfied. However, only 0.76% of the total observations were from shared rooms, meaning that it is possible our data is not very reliable and therefore leading to bad results in the model.

Future Work

As for future work we would like to duplicate this study but in cities in the United States to determine if the same methods to increasing guest satisfaction would apply. This would be interesting as US cities tend to be more spread out than European cities and have worse public transit meaning that it would be even more important to be near the city center. Also, we would like to research further into the relationship between price and guest satisfaction as there is a statistically significant correlation between the two, however, the correlation that a higher priced room would be more likely to leave a guest satisfied, while holding all other variables constant, is interesting but not very intuitive. It would be interesting to run an experiment where we listed identical Airbnbs for different prices to determine whether or not this relationship holds true.