Predicting the Likelihood of an Airbnb Being Booked

Eeshaan Ashok

Feature Engineering

The first step that I took in creating this prediction model was adding features to the dataset that could differentiate myself from other datasets. First, I created some basic metrics from features that were present in the previous dataset. I made the duration of service in months that I got by taking the number of reviews and dividing it by the number of reviews per month. This way I can see how long an airbnb has been in service. This way I can try to see if an airbnb has been in service longer it could have better ratings. Also, I made a price per minimum nights so I could see on a rate basis how much an airbnb would be per night. The next thing I tried to find was price competitiveness. For this categorical feature I found the average price of an airbnb in each neighborhood in the city and if the price of an individual airbnb was more than the neighborhood average it would be not competitive and if it was less than the average price it would be marked as competitive. If it is equal it is marked as equal. Obviously, this is not an all encompassing metric. This is just taking price in a vacuum and not considering anything else. So it may be a little flawed but is it something that I thought could be useful.

Something that I tried to incorporate was trying to merge my dataset with a restaurant data set. I wanted to try to see how many restaurants would be in the vicinity of each airbnb. To do this I found a restaurant csv and used latitudes and longitudes of airbnbs and restaurants to try to calculate distances between the two and return the amount of restaurants that were within .5 miles of an airbnb. However, both of these dataset are quite large, and the amount of time that it would take for this code to run did not make sense, so I decided to scrap the idea.

The next thing I pursued was something with subway station data. I found a subway station data set that had latitudes and longitudes of every subway station in New York City. Then similar to the restaurants dataset I found all subway stations that were within 0.5 miles of an

airbnb. The thinking behind this is that all subway stations do not have all the trains so by being near more subway stations there are more places in the city that are easily accessible to a person. For example, if someone is only near 1 subway station and that subway only has the 1, 2, 3, and L trains. The person has to go through more trains to get to certain places in the city, and for a tourist this could be very confusing. By being near 5-10 stations it would almost guarantee that all the subway lines would be running within walking distance from them so they can minimize the amount of subways and confusion that tourists would experience on a trip.

I also brought over features that I used in the previous part that used the attractions csv, such as distance to nearest attraction.

These are all of the features that I engineered that I thought would help my model predict the likelihood that an airbnb would be predicted.

Model Creation

Now that I had some new features to use I started building my model. The model was going to predict the likelihood that an airbnb will be booked. This is different from deal quality because a deal could be good quality but it may not get booked for a number of reasons. Also, now I had more features that could justify this based on travel.

There are many conditions for an airbnb to be highly likely to be booked, since I wanted to make sure that it was complex. The highlights of the features are a high number of reviews, and reviews per month. This shows that many people are booking the airbnb already so this will likely continue. Also, short minimum nights was important because it shows that properties are more flexible with their booking policies since it could attract more guests. Properties with a price below 200 a night and high mean review scores above four are selected. This is self explanatory. Good prices and ratings speak for themselves. Good deal quality was stolen from

the previous assignment since I felt that it was a good all encompassing indicator that could be used. Then two things that I really wanted to emphasize are distance to attractions and the number of subway stations that were near the apartment. The criteria I used was if the attractions are within 500 units of the airbnb and if there are 5 or more subway stations near the airbnb. This means that they will have easy accessibility to most of the city through the subway, and they are also near a tourist attraction so the location would be more desirable. Another thing that was emphasized is the borough being Manhattan. I think that most places in Manhattan would be booked regardless of most of these features since they already meet the previous criterias. So I made sure my model took this into consideration and placed Manhattan bookings more likely in the highly likely to be booked category.

Then for likely to be booked, the categories were quite similar to the ones in the highly likely category, but they are in less desirable boroughs. So it includes all the other boroughs except for Manhattan. These are quality bookings, it's just that they aren't in the main part of the city and it is reflected in the category that it was placed in.

The unlikely category is if they aren't fitting in any of these categories. Places with low ratings, longer minimum nights, and less convenient locations would fit into this category. These could be booked but they wouldn't be the top recommended airbnbs by any means. If they don't fit anything then it is placed in the other category. I made sure that the other was very rare and more likely it is placed into the unlikely category.

Possible Student Approach to Predicting Likelihood Categories

A student investigating this data set would need to start by plotting various features against the likelihood ratings so find correlations. They could make scatter plots and histograms to find this. This could help them visualize correlations and it could help them prioritize the

features that they want to use in their prediction. They can try to plot things like a mosaic plot to see a possible correlation between borough and booking likelihood. They can also try to plot a box plot to compare the number of subway stations near an airbnb and booking likelihood. By doing this the students would be on the right track to finding the pattern as these would are the most important features. They would also need to recreate the features that I engineered since they are deeply ingrained in my model and I would give them the subway csv for them to use as well. Without these features it would be difficult to predict the booking likelihood. However, I have incorporated some noise in the data set like host tenure, which has no effect on the model and it could cause students to include unnecessary features in their model, but by performing exploratory analysis and creating correlation plots this can easily be found out and ignored.

The student can then move on to experimenting with different machine learning algorithms to try and predict the booking likelihood. They may start with simpler models such as logistic regression or decision trees and slowly move onto more complex models like random forest models or even gradient boosting machines. During the training, they may use methods such as cross validation and using calculated methods like accuracy in order to see how close they are to the actual data. This could help them test the performance of different models and select the one with the best overall performance. During this step they may also experiment with feature selected techniques to help identify the most important features for predictions.

By doing this students can explore the data set in order to find the pattern that would lead them to an accurate prediction model.