Aiden Conklin

Final Project Written Report

Clustering:

A tibble: 6 × 6

clusterGroup <int>	accommodates <dbl>	bathrooms <dbl>	bedrooms <dbl>	beds <dbl>	price <dbl>
1	2	1	0	1	102.0
2	5	2	2	2	156.0
3	4	1	1	1	136.5
4	2	1	1	1	78.0
5	4	1	1	2	115.0
6	8	2	3	4	259.0

6 rows

- This table shows each cluster group and its notable attributes.
- This can be broken down from each cluster by their distinct characteristics.
    - Cluster 1 can be an affordable studio apartment for a group of 2 people.
    - Cluster 2 can be described as a family retreat that is comfortable for the average family
    - Cluster 3 might be a little more pricey for a small group or couple that want a little more space in a nice area.
    - Cluster 4 is the cheapest option for people not planning to spend time in the Airbnb (think backpackers or budget travelers).
    - Cluster 5 is a more compact room that may be tight but has living space for everyone at a cheaper price.
    - Cluster 6 is the largest Airbnb, and it is best for large groups while being the priciest.

This clustering could be very useful for Airbnb. For example, each cluster can become a category that users use to filter out Airbnbs that they may not have a budget for or are not interested in. These clusters provide an automated way to filter each Airbnb in a way that makes sense.

Price Prediction:

- For my final prediction, I used a random forest model.
- My model was fairly accurate ending with an average error of $73.64 for each Airbnb listing in the Kaggle submission.
- I think that moving, forward with the model I created can be appropriate; however, with more time and data tweaking, I do think a better model can be made.
- The two most important pieces of data when determining price is consistently bedrooms and bathrooms. As these two factors go up, the price also trends upwards.
- In relation to the host, as both the host's number of reviews and total listings go up, the price tends to trend downward.
- I think that if we gathered more information on the surrounding area we can make our predictions more accurate. For example, total restaurants in a specific radius or maybe if it's in a neighborhood or urban area.

Technical Report

Clustering:

- Before clustering, I cleaned the data by imputing medians for NA values, logging numeric values, and standardizing these values.
- I excluded non-numeric values as clustering does not work as well with nominal values without extra encoding.
- I used k means to make the clusters.
- I chose the number of clusters using an elbow plot of the total within-cluster sum of squares

Price Prediction:

- To clean the data I normalized it, and got rid of highly correlated predictors, and near-zero variance predictors.
- I also mutated the description to use a count of characters instead of the actual description to make it more useful.
- Before modeling, I excluded variables that wouldn't provide any information(such as IDs, URLs, and names)
- The 3 models I considered were Decision Trees, K Nearest Neighbors, and Random Forests. I chose Random Forests as I felt it was able to encapsulate the data best and was the most accurate.
- **Decision Trees** had an RMSE or accuracy of 100.19 in the trained data.
- **KNN** had an accuracy of 86.48.
- Finally, The **Random Forests** model had an accuracy of 40.11.
- For the parameters in Random Forests…
    - **MTRY** tuned with a range of (2, 6) which affects the number of predictors tried at each split
    - **MIN_N** tuned with a range of (2, 10) this affects the minimum number of observations per leaf.
    - **TREES** I used 500, this represents the number of trees used in the forest, I felt this number was an appropriate amount to cover all bases without overfitting.