Airbnb: Effect of Neighborhood Factors on the Occupancy Rate

Anushital Siddiqi (asiddiqi30), *Fei Chi (*fchi6), *Gaurav Chakravarty* (gchakravarty6), *Mohammed Kashif Uddin Ansari* (mansari44), *Saikrishnan S*ubramanian (ssubramanian99), *Sharat Chandra* (schandra46)

---

1. Introduction

Established in 2008, Airbnb has become quite a phenomenon in the last few years with over 3 million listings in over 191 countries and over 150 million guests [7]. It has led to financialization of houses by creating a hyper-flexible rental market [10]. This has resulted in gentrification of neighborhoods introducing a new potential revenue flow into housing markets [9], creating micro-entrepreneurship opportunities that empower individual landlords [6], many of whom are now moving their properties from long term leases to short term lets on Airbnb and in some case putting their idle space on short term rent [4]. Simultaneously, the profile of Airbnb hosts is also changing from regular landlords to professional investors. Consequently, there is an increase in demand for decision-making tools for Airbnb investments. An example of such a decision is, "What property and neighborhood attributes are best for Airbnb short-term lets?" For instance, should they invest in a good property in an average neighborhood or an average property in a good neighborhood. This is the inspiration behind our research project: to answer a subset of these difficult questions that an investor face.

2. Problem Definition

The objective of our project is to explore the impact of neighborhood factors namely - tourist attractions, restaurants, and median household income on the occupancy rate of Airbnb rentals. We will control external variables to isolate their effect to the extent possible. This analysis along with the visualizations will be useful input for individuals or companies looking to invest in Airbnb properties.

3. Survey

There are many interesting studies on Airbnb dataset like predicting occupancy rate using host controlled factors [2], using utility based factors to estimate consumer's willingness to pay for Airbnb rentals [11]. One of the papers developed classifiers to predict the price and the neighborhood of the property. Predicting a listing's neighborhood gives us insight into the cultural elements visible through text, image, and amenities that might link neighborhoods together, and could potentially be applied for a recommendation system (e.g. "If you enjoyed your stay in the Haight Ashbury, we recommend trying Alamo Square!")and predicting the price of a listing can be helpful in developing internal pricing tools that Airbnb can offer to its hosts[12].

However, studies on the impact of neighborhood factors (restaurants, tourist attractions and median household income, access to public transportation and other amenities etc.) on Occupancy Rate of Airbnb rentals seem to be missing. This project aims to start filling this gap.

Studies predicting Prices or Occupancy Rates have used Regression models, mainly LASSO, Ridge, Elastic Net [13]. Tibshirani in one of his papers discusses how LASSO produces interpretable models like subset selection and exhibits the stability of Ridge Regression and how the former can be applied in a variety of models like generalized regression models and tree based models [15].

One paper claims that XGBoost and random forest demonstrated the strongest performance in predicting rental prices with the test data collected from the Airbnb website on January 28th, 2020 [8].

This inspired us to use XGBoost for our project too because of its many benefits as discussed later in the paper.

Sentiment Analysis is another area with extensive research on the Airbnb data because reviews have a profound effect on which places to visit or book like restaurants, tourist attractions, or an Airbnb rental [17]. Airbnb prices are also influenced by the review score, the characteristics of the room, and the features of the neighborhood [18]. Sentiment Analysis also helps in improving the service and rental quality by giving the hosts an insight into the guests' requirements [14].The sentiments derived from reviews can be classified as negative, neutral or positive . It has been observed that using classified sentiments with a Regression Tree provides a more accurate prediction result compared with using numerical sentiment scores [16]. Although sentiment analysis is an interesting approach, we will not be using it as it may be too complex to implement within the time frame allocated for this project and our chosen algorithms seem to give a desirable performance with the predictors we have selected.

4. Proposed Method

4.1 Intuition

Generally, investors focus on internal factors like the property & service quality and other host control factors before investing in an Airbnb property and they often ignore a crucial aspect - the neighborhood factors. Since these factors are important for the renters, ideally, they should have an effect on the occupancy rate and thus on the profitability of the property. We wanted to find this relationship and include the neighborhood factors in investing decisions to determine if they have an effect on the occupancy rate.

4.2 Algorithms

Several approaches could be taken for this research study. We first ran a regression algorithm - LASSO, assuming there was a linear relationship between the predictors and the response variable. We chose LASSO because it handles feature selection automatically. But it gave us a lower value R- squared when compared to the same values obtained from XGBoost. We will discuss this later in the report. We can see in the correlation heatmap plot in Figure 1 that all of the predictors have weak correlation (below 0.4) with availability_365, from which we have derived the occupancy rate explained later in the report. Low correlation and low R-squared value are indicative of the fact that the factors and the response variable may not have a linear relationship. This led us to the next best model: **XGBoost**, which is short for **Extreme Gradient Boosting.** The two primary reasons for using XGboost in our design are high execution speed and model performance. Gradient Boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, mainly decision trees. To find a weak learner, it applies base learning algorithms with a different distribution. Each time a base learning algorithm is applied, it generates a new weak prediction rule. This is an iterative process. After many iterations, the gradient boosting algorithm combines weak rules into a single strong prediction rule. The Gradient Boosting algorithm mainly has three elements: a loss function to be optimized, such as cross entropy for classification or mean squared error for regression problems, a weak learner to make predictions, such as a greedily constructed decision tree and an addictive model, used to add weak learners to minimize the loss function.

4.3 Visualization

We have put 7 different interactive visualizations on Tableau Public. 3 of them are maps which take zip code from the user and return the following: restaurant count by property, attraction count by property, occupancy rate by property, 1 of them is a map which takes desired occupancy rate from the user and returns the zip codes with that occupancy rate. Additionally we have 3 different line charts: occupancy rate vs nearby places (restaurant or tourist attraction) by bedroom, occupancy rate vs nearby places(restaurant or tourist attraction) by room type, occupancy rate vs nearby places(restaurant or tourist attraction) by property type.

Of all these visuals, we found the one in Figure 3: Occupancy rate vs Restaurant Count for 1 bedroom, 2 bedrooms, 8 bedrooms to be most interesting. 1 bedroom and 2 bedrooms do not show a particular pattern in the occupancy rate but 8 bedrooms does. The occupancy rate for 8 bedrooms jumps sharply beyond 18 nearby restaurants. This could probably be because the properties with 8 bedrooms are hostels or small hotels. So investors should keep in mind that if they are buying a big property, they should ensure that it is located in a busy area with lots of restaurants, cafes, eating joints etc. to ensure a high occupancy rate.

4.4 User Interface

We have created the GUI using Flask. The GUI accepts user inputs related to an Airbnb property as shown in Fig 2., and returns the predicted occupancy rate by using the trained model and using it for generating the occupancy rate.

5. Experiments and Evaluation

5.1 Data

We downloaded the Airbnb data from Insideairbnb.com for Austin City, Texas. This data is 7.2 MB on disk and has approximately 10,000 rows. Additionally, we also downloaded data from census.gov to get the median household income for every zip code used in our study. This data is 182kb on disk.

To get the data on the number of nearby restaurants and tourist attractions of an airbnb property, we called the Google Places API. The size of the Google Places data is very large and it took significant time to collect relevant data for our project. Since each row of our dataset is a property, we called 2 APIs for each row - one for the number of nearby restaurants and the other for the number of nearby tourist attractions. Each API was roughly 100 KB. The total number of API calls made were 2 * 10,000. The size on disk of our raw input data(gathered from google) is 100* 2* 10,000 = 2000000 KB or 2 GB approximately.

5.2 Features

Since LASSO does its own feature selection, we didn't have to do feature selection. However, in the case of XGBoost, we applied the best practices of feature extraction. For the column bedrooms, we filtered out the number of bedrooms above 9 because their frequency was quite low as they appeared in a total of 19 rows. The column zip code has 40 unique values but we don’t have enough data to model at that level of granularity. So we moved from 5 digits zip code level to 4 digits level which reduced the number of zip codes from 40 to 6 where the latter was more representative of the data and was also manageable while being one hot encoded. The column amenities had approximately 1000 unique

values. We did iterations to choose 7 amenities which had the most impact on accuracy rate of predictions and then one hot encoded them.
We dealt with the null values of the three neighborhood factors differently. We replaced the null values of the median household income by the median value of that column, the null values of the restaurant count and the tourist attraction count by zero and we dropped the null values of the zip code column altogether.
Since the occupancy rate is not directly provided in the Airbnb dataset, we derived it from the column availability_365 using the following formula:
Occupancy rate = **1 - availability_365 / 365**

5.3 Experiments
We used LASSO and XGBoost to run the following experiments:
I. Study the effect of the number of nearby restaurants on the Occupancy Rate.
II. Study the effect of the number of tourist attractions on the Occupancy Rate.
III. Study the effect of median household salary of a neighborhood on the Occupancy Rate.
IV. Study the combined effect of the following neighborhood factors on the Occupancy Rate: median household income of a neighborhood with the number of nearby restaurants, median household income with the number of tourist attractions in the neighborhood, and finally the number of nearby restaurants with the number of tourist attractions in the neighborhood.
V. Study the effect of all the above features combined on the Occupancy Rate of the Airbnb Property.

5.4 Evaluation
As discussed earlier in the Algorithms section, we ran LASSO and XGBoost to study the effect of our three chosen neighborhood factors on the occupancy rate of the Airbnb rentals. For both the models, we divided the 70% of the dataset into a train set and 30% into a test set. To improve the models' performance, we tuned the hyperparameters, using Sklearn GridSearchCV and conducted a 5 foldCross Validation. On comparing the two models, we concluded that XGBoost is a more suitable model for our study.
We compared the Adjusted R-Squared values of both the algorithms to choose the better of the two. Adjusted R-Squared value is an indicator of how much of the variability in the data is explained by the given model. The LASSO model applied on our data has an Adjusted R-Squared value of 0.217 on the test set. This means that this model is able to explain only 21.7% of the variability in data, which is quite low when compared to XGBoost, which has an Adjusted R-squared value of 0.4614. So, we selected XGBoost to conduct our study since it explains the variability in our data significantly better than the LASSO model.

6. Innovation
Our study has the following novel contributions: using three different data sources, as explained in the Experiments section, to study the effect of neighborhood factors - number of nearby tourist attractions and restaurants and median household income on occupancy rate, controlling for the effect of Airbnb features that we get from the listings.csv file of the Airbnb dataset and provide a tool for prospective Airbnb investors to help determine the effect of neighborhood factors on the occupancy rate and accordingly include this factor while making Airbnb investing decision. Other researchers have explored

the connection between spatial location of Airbnb properties to hotels [1], the effect of tourism clusters on Airbnb performance [2] and measurement of the effect of factors in control of the host to Airbnb occupancy rate [3]. We are building upon these studies and improving them by combining 3 important neighborhood factors to explore their impact on the occupancy rate of Airbnb rentals. To our knowledge, this is the first attempt to combine these three important neighborhood effects with a lens to investor decisions.

7. Plan of Activities

All team members contributed equally to the project except Sharat Chandra. He didn't finish any of the tasks assigned to him and other members had to pick up the slack causing unnecessary stress to the entire team.

8. Conclusion and discussion

The adjusted R-Squared value (test set) for all predictors without the neighborhood factors is 0.46142. This is our benchmark value against which we compared our neighborhood factors individually, in combination of twos and all together. When we include all the 3 neighborhood factors together, adjusted R-Squared increases to 0.48565. There is clearly an increase in the explaining power of the model, which means that the neighborhood factors together definitely have a positive impact on the occupancy rate. All the neighborhood factors when added individually also have an impact on the occupancy rate because all of them increase the adjusted R-Squared value. However, median household income stands out as the most impactful neighborhood factor because when added individually, it has the highest adjusted R-Squared value of 0.48023 as compared to 0.46733 for restaurant count and 0.46765 for tourist attraction count. Even when we compare the combined effect of two of these factors at a time, the combinations of median income perform better than the other two factors. Based on our analysis, we can conclude that the neighborhood factors have a positive impact on the occupancy rate and median household income has the strongest effect. So, investors who want higher occupancy should invest in Airbnb properties in richer areas. And based on the visualization in Figure 3, we can say that if investors want to invest in bigger sized properties, they should invest in busy areas like downtown which have a higher number of restaurants, cafes etc.

Going forward, further research can be conducted to discover other powerful neighborhood factors that affect the occupancy rate of Airbnb properties like neighborhood walkability score, neighborhood crime rate, neighborhood gentrification levels, distance from concert venues, distance from sporting venues, distance from national parks, distance from grocery stores etc.

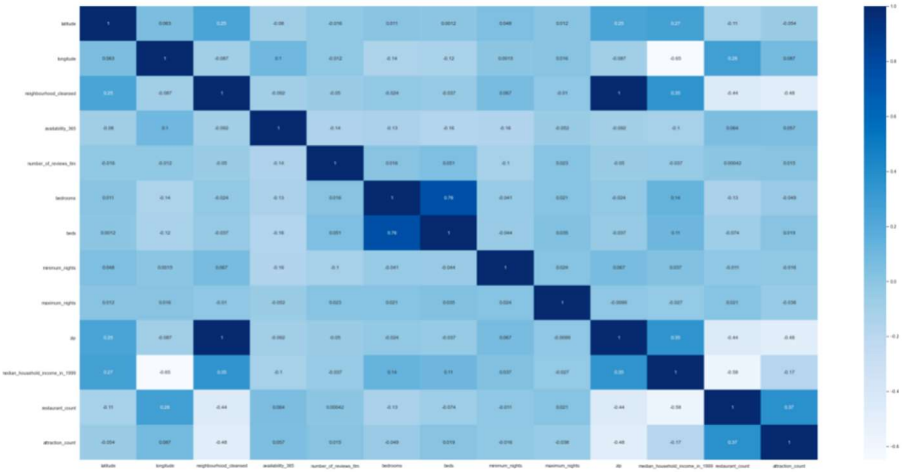

Figure 1: Correlation HeatMap

Figure 2: Interactive User Interface

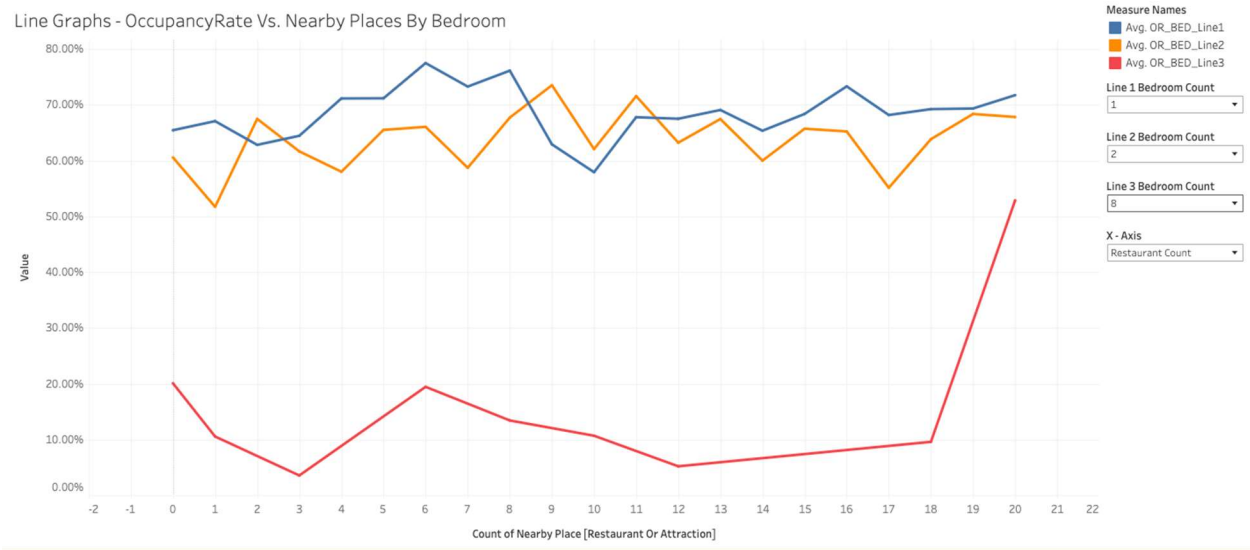

Figure 3: Occupancy rate vs Restaurant Count for 1 bedroom, 2 bedrooms, 8 bedrooms

References:

1. Gutierrez, J. , Garcia-Palomares, J. C. , Romanillos, G. , & Salas-Olmedo, M. H. . (2016). Airbnb in tourist cities: comparing spatial patterns of hotels and peer-to-peer accommodation. Tourism Management, 62, 278–291.https://arxiv.org/pdf/1606.07138.pdf
2. Lee, Y.-J. A., Jang, S., & Kim, J. (2020). Tourism clusters and peer-to-peer accommodation. Annals of Tourism Research, 83, 102960. https://doi.org/10.1016/j.annals.2020.102960
3. Mendoza, J.V., Aquino, J.M., Briones, K., Geralde, J.M., Macasaet, J.M., Mirambil, A.R., & Meñez, M.F. (2019). Determinant Factors of Airbnb Occupancy Rate in the Province of Batangas, Philippines.
4. Koster, H., van Ommeren, J. and Volkhausen, N., 2021. *Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles*. [online] https://www.sciencedirect.com. Available at: <https://www.sciencedirect.com/science/article/pii/S0094119021000383> [Accessed 7 November 2021].
5. Krajcik, V., Kljucnikov, A., Rihova, Elena.(2018), Prediction of Occupancy Rates Based on the known parameters for different types of housing, The 12th International Days of Statistics and Economics, Prague (LASSO Regression for occupancy rate)
6. Mahmuda, S., Sigler, T., Corcoran, J., & Knight, E. (2021). Airbnb and micro-entrepreneurship in regional economies: Lessons from Australia. Geographical Research, 1– 17. https://doi.org/10.1111/1745-5871.12506
7. Martinez, R.D., Carrington, A., Kuo, T., Tarhuni, & Abdel-Motaal, N.A.Z.,The impact of an Airbnb host's listing description 'Sentiment' and length on occupancy rates. (n.d.). arXiv.org. https://arxiv.org/abs/1711.09196
8. Koster, H., van Ommeren, J. and Volkhausen, N., 2021. Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles. [online] https://www.sciencedirect.com/. Available at: <https://www.sciencedirect.com/science/article/pii/S0094119021000383> [Accessed 7 November 2021].
9. Wachsmuth, D. and Weisler, A., 2021. Airbnb and the rent gap: Gentrification through the sharing economy - David Wachsmuth, Alexander Weisler, 2018. [online] SAGE Journals. Available at: <https://doi.org/10.1177/0308518x18778038> [Accessed 7 November 2021].
10. Cocola-Gant, A., & Gago, A. (2019). Airbnb, buy-to-let investment and tourism-driven displacement: A case study in Lisbon. Environment and Planning A: Economy and Space, 53(7), 1671-1688. https://doi.org/10.1177/0308518x19869012
11. Chen, Y., & Xie, K. (2017). Consumer valuation of Airbnb listings: a hedonic pricing approach. International Journal of Contemporary Hospitality Management, 29(9), 2405–2424. https://doi.org/10.1108/IJCHM-10-2016-0606
12. Tang, E.; Sangani, K. Neighborhood and Price Prediction for San Francisco Airbnb Listings. Available online: http://cs229.stanford.edu/proj2015/236_report.pdf
13. Rezazadeh Kalehbasti, P., Nikolenko, L., & Rezaei, H. (2021). Airbnb price prediction using machine learning and sentiment analysis. Lecture Notes in Computer Science, 173-184. https://doi.org/10.1007/978-3-030-84060-0_11

14. LEE, Y., 2021. Investigating Citizen Perceptions and Business Performance of Airbnb in Korea. [online] Koreascience.or.kr. Available at: <https://www.koreascience.or.kr/article/JAKO202120953722354.page> [Accessed 7 November 2021].
15. R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.
16. Liu, P. (2021). Airbnb Price Prediction with Sentiment Classification. Master's Projects, (979). Retrieved from https://scholarworks.sjsu.edu/etd_projects/979.
17. Agüero-Torales, M. M., Cobo, M. J., Herrera-Viedma, E., & López-Herrera, A. G. (2019, December 31). *A cloud-based tool for sentiment analysis in reviews about restaurants on TripAdvisor*. Procedia Computer Science. Retrieved October 14, 2021, from https://www.sciencedirect.com/science/article/pii/S1877050919320125.
18. Lawani, A., Reed, M., Mark, T. and Zheng, Y., 2021. Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston. [online] www.sciencedirect.com. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S016604621730340X> [Accessed 7 November 2021].

Websites:

1. https://support.sas.com/rnd/app/stat/papers/2015/PenalizedRegression_LinearModels.pdf
2. https://en.wikipedia.org/wiki/XGBoost