University of Essex

Department of Computing

Machine Learning 2024

Development Team Project: Project Report

Group 1

Date: 10/05/2024

Word Count: 1023

1. Introduction

Airbnb is an online marketplace that allows individuals to either host a place for travelers to stay or book a place to stay with a host. Over the past several years, millions of tourists have used the service (Guttentag et al., 2017). But what exactly influences a traveler's choice in where they stay? Knowing why a traveler books a listing is important for hosts to optimize their airbnb rental unit, leading to greater revenue for Airbnb and the host, as well as increased satisfaction for all sides: the traveler, the host, and Airbnb themselves. It may also encourage potential hosts that have rental units that fit the ideal standards for travelers to increase the availability of the unit, leading to more traffic for Airbnb's services. Using data from the listing activity and metrics in New York City for 2019, this report attempts to answer the following: how do various factors influence the likelihood of a listing being booked?

2. Exploratory Data Analysis & Data Pre-processing

Raw experimental data is highly susceptible to noise, missing values, and inconsistencies, making initial data cleaning a critical step before any meaningful analysis can occur (Calabrese 2018). Our first task involved exploring and cleaning the dataset to ensure it was suitable for addressing our business question. As illustrated in Figure 1 of the code, we began by identifying the variables and rows within the dataset. Utilizing the pandas library, we found that the dataset comprised 16 variables and 48,895 entries. To delve deeper into the variables, we employed the describe command to generate summary statistics (Figure 2), providing us with an overview of the data's central tendencies and variability. Additionally, we used the 'isnull' command to pinpoint missing values, discovering that four variables contained null entries. To address these missing values, we either removed the affected rows or replaced the missing values with 0, as shown in Figure 3. This thorough data cleaning process was indispensable for ensuring the accuracy and reliability of our subsequent analysis and findings.

Figures 4 through 7 present various visualizations that provide insights into the dataset's composition. These figures illustrate the count of different room types, the distribution of listings across neighborhoods and neighborhood groups, and the number of room reviews. Notably, 'entire home/apartment' emerges as the most frequently listed room type, with over 600,000 reviews. Manhattan stands out as the area with the highest number of listings, totaling 20,000, with Harlem contributing 2,500 of these listings. Furthermore, Figure 8 highlights the median price analysis, revealing that Manhattan is the most expensive neighborhood group, with an average booking cost exceeding $140. These visualizations not only help in understanding the dataset's structure but also provide a foundation for more detailed data analysis and strategic decision-making based on the identified trends and patterns.

3. Machine learning

Machine learning (ML) is an integral part of modern data science, allowing meaningful insights and predictions to be made from vast amounts of data. In order to create a ML

algorithm, first the normalizer from the sklearn.preprocessing module was used to scale the features by adjusting each feature vector to have a length of one. This ensured that each data point contributed equally to the model and that features with larger magnitudes didn't have larger impacts on the future model and analysis (Verma (2023). After preprocessing, the data was split into training and testing sets. This split allows the model to be trained on one subset of the data and evaluated on another, ensuring that the model's performance can be generalized to new, unseen data (Bishop, 2006). Using this process, we reassured that the model does not overfit the training data, increasing the reliability of its predictions.

In the context of clustering - an unsupervised learning technique - K-Means clustering is applied using KMeans from sklearn.cluster. This algorithm partitions the data into k clusters, where each data point belongs to the cluster with the nearest mean (MacQueen, 1967). Figures 13 and 14 display the output of the initial K-Means clustering algorithm, where the groups show the booking likelihood based on the location of the listing. After the algorithm was created, a test was done to verify which K-value, or the number of groups that can be created without overfitting, would be best (Figure 15). After updating the algorithm to reflect the new K-value, the group differences were more pronounced, with one group having a much higher booking likelihood than the other groups (Figures 16 and 17). Evaluating the quality of the clusters is done using the silhouette score, with higher silhouette scores indicating better-defined clusters (Rousseeuw, 1987). In this case, the silhouette score is 0.563, suggesting that the clusters are well-defined, distinct and the chosen number of clusters is appropriate for the data. K-Means was a good choice for our analysis due to its ability for simple implementation as well as being computationally efficient. By grouping the similar listings together, it allowed us to analyze the complex data and derive actionable insights. Other clustering methods such as Hierarchical clustering were considered but were opted against due to their impracticality with larger datasets, and their being computationally more expensive and less efficient.

4. Conclusion

Our analysis determined that several key factors significantly influence the likelihood of an Airbnb listing being booked. By examining data, we identified that price, number of reviews, location, and property type are crucial determinants of booking likelihood. Specifically, 'entire home/apartment' emerged as the most popular room type, and Manhattan was found to be both the most listed and the most expensive neighborhood group. For Airbnb, understanding these factors is critical for developing strategies that attract more hosts and guests, thereby boosting overall platform engagement and revenue. The insights from our analysis can guide Airbnb in tailoring their services and marketing efforts to align with market trends and traveler preferences. Future Airbnb hosts can leverage these insights to optimize their listings by strategically setting competitive prices, improving the quality and appeal of their property types, and accumulating positive reviews to enhance their booking likelihood. For travelers, the analysis provides a clearer understanding of what influences booking likelihood, helping them make better-informed decisions when choosing accommodations.

References

Guttentag, D.A., Smith, S.L.J., Potwarka, L.R., & Havitz, M.E. (2017) 'Why tourists choose Airbnb: A motivation-based segmentation study underpinned by innovation concepts', Journal of Travel Research, 56(3), pp. 396-409.

Calabrese, B. (2018) 'Data cleaning', in Ranganathan, S., Gribskov, M., Nakai, K., & Schönbach, C. (eds.) Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics. 1st edn. Oxford: Elsevier, pp. 472-481.

Bishop, C.M. (2006) Pattern recognition and machine learning. New York: Springer.

MacQueen, J.B. (1967) 'Some methods for classification and analysis of multivariate observations', in Le Cam, L.M. & Neyman, J. (eds.) Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Berkeley: University of California Press, Vol. 1, pp. 281-297.

Rousseeuw, P.J. (1987) 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', Journal of Computational and Applied Mathematics, 20, pp. 53-65.

Verma (2023) https://www.digitalocean.com/community/tutorials/normalize-data-in-python