**Project 1 Draft: Airbnb Nightly Rate Machine Learning Host Recommendation Modeling**

William Seth Justice

Bellevue University

DSC 680 Applied Data Science

Catherine Williams

July 5, 2021

**Project 1 Draft: Airbnb Nightly Rate Host Machine Learning Recommendation Modeling**

Airbnb is perhaps one of the most well-known online marketplaces for lodging accommodations across the globe. The Airbnb online market place connects hosts with those looking for stays through a mobile app or website. In 2020 Airbnb had a revenue of just over $3.3B (USD). Airbnb hosts post their accommodation on the website or app and provide pictures, descriptions, and other details such as number of bathrooms, bedrooms, etc. The host then decides the nightly rate they charge based on their own market research. While the host should continue to be able to make their own decision regarding nightly rate, a recommendation model could be put into place which automatically evaluates similar accommodations and offers a nightly rate suggestion based on similar stays in the area. Airbnb charges a 3% fee on all bookings. Most hosts pay a service fee of 3%, but Airbnb Plus hosts, hosts with listings in Italy, and hosts who use Super Strict cancellation policies may pay more. This fee is calculated from the booking subtotal (the nightly rate plus cleaning fee and additional guest fee, if applicable, but excluding Airbnb fees and taxes) and is automatically deducted from the host payout (Airbnb, 2021).There is a potential that hosts either overpricing or underpricing their accommodation could create missed income due to decreased earnings or even customers turning to alternative bookings such as traditional hotels. If a recommendation model is correctly implemented this could offer the host the most up to date market research available to quickly propose the most competitive rate aligned with similar stays in the area. Additionally, this recommendation would minimize the potential for hosts either over or under pricing their accommodation thus minimizing missed earning potential to both hosts and Airbnb.

**Exploratory Data Analysis**

The dataset chosen to perform this was collected from Kaggle and was compromised of Airbnb bookings and customer feedback in Seattle Washington dated from 2008 to 2019. There were around eighty thousand observations. At an average rate of just over $127 per night that equates to an annual company income of just under twenty eight thousand dollars per year in this single market.

There are a few important indicators that are important in this dataset that are likely similar in almost every geographic area as well. The first being the customers gearing towards a private dwelling of their own. Not just a private room, but a fully private accommodation without hosts or other travelers in the same booking. This can be seen in the visualization below.

The next important consideration is proximity to other locations such as downtown areas, airports, and other attractions. Since these locations are where customers want to book more often, they are willing to pay a higher rate to be closer to such attractions. The two visualizations below are meant to show the correlation between specific geographical areas, in this case zip codes, and higher prices and increased bookings. It can be seen in the map that there is a central location (Downtown Seattle/Seattle Proper) that is most sought after by customers.

**Feature Reduction and Feature Engineering**

After joining the correct datasets the first step was to conduct feature engineering and feature reduction. In attempts to decrease model error the first engineered feature was created utilizing the NLTK VADER sentiment analysis package. The purpose of this package in regards to this project is to add a quantitative figure derived from the unstructured customer comments. Valence Aware Dictionary for sEntiment Reasoning, or Vader, is a NLP algorithm that blended a sentiment lexicon approach as well as grammatical rules and syntactical conventions for expressing sentiment polarity and intensity (Ma, 2020))What was noticed in this task was that the sentiment analysis final figure was likely not sensitive enough to add much weight to the model. The compound score figures are valued between negative one and positive one, with the closeness to positive one being a positive sentiment and closeness to negative one being a negative sentiment. An example of this problem is the fact that over half of the observations fell above the 0.9 marker. Even more important was that there were only five observations of “negative” sentiment comments that fell below the zero score. This tells us that the overall sentiment of customer feedback is extremely high while using the NLTK VADER sentiment analysis tool.

The next feature engineering step was to derive some sort of quality factor to the listing. In this sense quality refers to the “updatedness” of the booking. In order to derive this a formula was established to create a score from negative one to a positive one. This was also done using the NLTK VADER sentiment analysis tool. The sentiment analysis in this case was applied to the host’s description of the booking. It should be noted that given the data available there is little ability to derive an insightful “updatedness” quality score. This will be further discussed in the final paragraphs of this report.

The final feature engineering task was to derive a score based on the popularity of the location. This is known as the Zip Code Popularity Score. The intended result here is to create a feature that encompasses the distance to local attractions in the area. This zip code popularity score was simply a ranking structure of zip codes by the count of the number of observations with one being the highest and thirty being the lowest. After consideration, it has been determined that using the geolocation coordinates and applying them to determine a number of miles to downtown or other popular attractions may be better suited. Although, this approach may be insufficient when attractions are located outside this area. For example, a large festival which attracts many booking that is not located in the downtown area. This also changes from city to city and would need to be determined for each market area.

The next step was to conduct feature reduction. The dataset after feature engineering had over ninety features. Many of these features could be easily excluded as they are primarily a byproduct of the web scraping process. Examples of these include variables such as host url, booking url, host picture, scrape id, etc. After removal of these came actual feature reduction by means of ANOVA testing to determine statistical significance to the listed price. The ANOVA test in order of statistical significance resulted in bathrooms, room type, bedrooms, cleaning fee, accommodates, reviews per month, review check in score, guests included, review scores rating, zip code, review scores value, neighborhood group, property type, host response time, host compound score, and customer compound score.. What was not found to be statistically significant were the engineered features, but they were left in the model to encompass the project challenges.

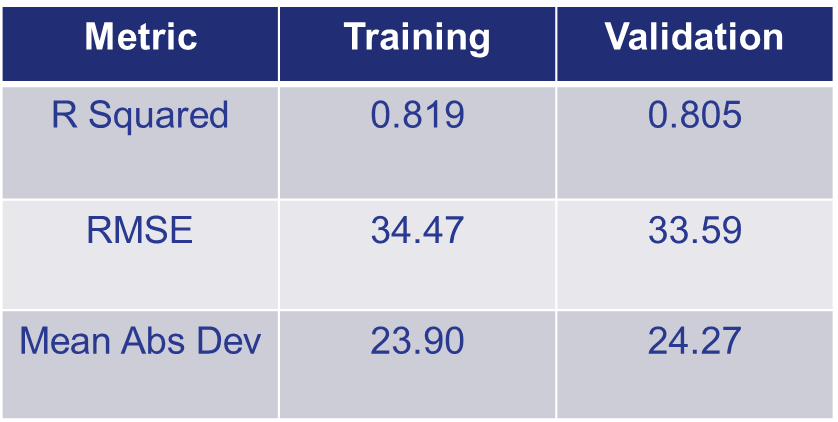
**Outlier Handling**

The dataset contained a fair amount of outliers. Since this data was scraped from the Airbnb website it can be concluded that these outliers are valid but not valuable. Some examples of these include extremely high prices and fees. With the overwhelming majority of nightly rates being under the three hundred dollar range, there were nine outliers that existed above the seven hundred fifty dollar mark. It can be assumed the models will be sensitive to this and they will increase error.

**Modeling**

Three models were compared for predicting the recommended nightly rates. The models chosen were a neural network and a regression model. The metrics of performance for model evaluation are R-squared, and root mean squared error (RMSE). The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model’s predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction (Grace- Martin, et al., 2020).

The neural network model performed best. The R squared value for training was 0.81 and 0.80 for validation. RMSE was $34.47 for training and $33.59 for validation. So the validation runs out performed the training run.



The regression model performed the second best. The R squared was 0.69 for training and 0.67 with validation. Unfortunately, RMSE was $46 for both training and validation. It is great that the model performed extremely similar against training and validation but an error figure of just under fifty dollars is too high for deployment to hosts.

**Recommendations and Further Considerations**

After modeling this challenge it has been determined that the neural network may perform well enough for deployment. The linear regression model did not. It must be agreed upon at a specified dollar amount threshold for deployment.

The predictive analytics team has two recommendations for improving model performance and they come by way of data gathering. The first data that needs gathering is that of reliable booking quality defined above as “updatedness”. It is clear that the engineered data did not fit the bill as statistically significant. The second data gathering would be that of a score to provide a walking distance to area attractions. It is not clear if this will be a significant feature in modeling but could be valuable for both the researching consumer as well as the predictive analytics team.

Utilizing the SMART business methodology the outline of a way forward in these categories. SMART stands for **specific**, **measurable**, **attainable**, **relevant**, and **timely** (Decker, 2019). Specific: Gather additional data from customer reviews and research sentiment analysis tools with increased sensitivity. Measurable: Create a predictive model that is able to predict within a RMSE metric of $20 (USD). Achievable; TBD. The dollar amount of $20 USD may need to be adjusted and agreed upon with other departments. Relevant: This has the ability to improve income as well as potential to distribute to other companies for increased revenue. Time-bound: After a period of six months to gather data, it is the request to revist this challenge. Since much of the work is already completed and only required data implementation into existing models, a period of two weeks is anticipated for turn-around to presentation of final findings.

References

Airbnb. (2021). *What are Airbnb service fees? - Airbnb Help Center*. Airbnb. https://www.airbnb.com/help/article/1857/what-are-airbnb-service-fees.

Decker, J. (2019, December 18). *How to Write Your Business Plan, 2019 Style*. College of Business Administration. https://www.unomaha.edu/nebraska-business-development-center/news/2018/12/blog-business-plan-2019-style.php.

Grace-Martin, K., Ali, M. M., Kewal, Shawn, Ulrich, Mathur, N., Kris, Hashim, Dina, Yash, Sharma, K., Bohan, Noah, Brian, M., Fakhar, Syed, Murtaza, Jan, M. N., Kicab, … Grateful2U. (2020, November 2). *Assessing the Fit of Regression Models*. The Analysis Factor. https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/.

Ma, Y. (2020, February 5). *NLP: How does NLTK.Vader Calculate Sentiment?* Medium. https://medium.com/ro-data-team-blog/nlp-how-does-nltk-vader-calculate-sentiment-6c32d0f5046b.