AirBnb Rental Price Generator

Rakeen Rouf

Report for Airbnb Executives

Introduction

The goal of this analysis is to build an inference model to generate prices to aid Airbnb hosts (in Asheville, NC) to set prices for their short-term rentals. The model uses various explainable factors such as number of bedrooms, distance to downtown, desirable amenities, etc. along with advanced statistical techniques to fit a model. This model ultimately provides data-driven insights to hosts, thereby helping them optimize their rental income and/or competitiveness. The prices generated using this model should be considered as estimates of the average price a rental with similar characteristics should fetch (within acceptable bounds) in the ASHEVILLE, NC Airbnb market. By using this model, hosts can find out what makes their rental special and how to make it even better. They can focus on the things that guests value most, which can help them attract more bookings and earn more money.

The raw data set used in this analysis has 3,239 observations and 75 distinct columns.

Methods

This analysis uses a linear regression model to fit the data. Linear regression is a suitable choice for datasets with numerous variables and observations, which is the case with our Airbnb data set. Linear regression helps determine the level of influence each variable holds over the final outcome (e.g. rental price). By using this model, hosts gain insights into the expected pricing of their rentals. They can also gain a deeper understanding of what makes their listings unique and how to enhance their appeal. This empowers hosts to focus on the aspects most valued by guests, ultimately leading to increased bookings and revenue. `Following is a list of variables used to fit the model`:

`Number of bedrooms in the rental`: More bedrooms mean more guests can be accommodated, which can justify a higher price and vice-versa. The number of bedrooms is generally highly correlated with rental prices.

`Number of bathrooms`: Similar to bedrooms, the number of bathrooms is crucial for guest comfort and convenience. More bathrooms can lead to higher prices, as they enhance the overall quality of the rental.

`Type of room (Private room or Entire house)`: The type of room is a significant factor influencing pricing. An entire house typically commands a higher price compared to a private room, as guests have exclusive access to the entire space.

`Amenity score (Based off a customizable amenity scoring chart)`: Amenities play a crucial role in a guest's experience. Including an amenity score enables the model to quantitatively account for the value of different amenities, ensuring that listings with more desirable features are appropriately priced.

`Is the host a super host`: Super hosts often have a reputation for providing exceptional service and maintaining high-quality listings. This status can instill trust in potential guests, potentially justifying a slightly higher price.

`Host acceptance rate`: A high host acceptance rate reflects a host's willingness to accommodate guests, which can positively influence the perceived reliability and responsiveness of the host. This may be a factor in pricing.

`Distance to downtown`: Proximity to downtown or city centers is a key location-based factor. Listings closer to popular attractions or business districts tend to command higher prices due to the convenience they offer to guests.

Results and Conclusion

This developed linear regression model demonstrates a strong ability to predict Airbnb rental prices based on a variety of important factors. The adjusted R-squared value of 0.5468 indicates that approximately 54.68% of the variability in rental prices is accounted for by the included variables. The predictor coefficients reveal the impact of each variable on rental prices. Additionally, the high F-statistic of 294.3 signifies that the model as a whole is statistically significant, providing valuable insights for predicting rental prices accurately.

As an example of generating new prices, for a listing with 2 bedrooms, 1 bathroom, private room type, 8 total amenity score, 5 miles from downtown, 80 percent host acceptance rate and, no super host flag, the price would be $108.69. This price will be displayed along with certain recommended actions a host can perform to increase their prices/competitiveness.

With the model explaining majority of the variability in rental prices (even with the complexities involved), I am confident in recommending this for deployment. However, as with most model this model can also be further improved. Additional data such as seasonality, popular events in Asheville on the day of rentals and, macro trends (e.g. prices were lower during Covid-19) may help in drastically improving the efficacy of this model.

Report for Data Science Team

Introduction

The raw data set used in this analysis has 3,239 observations and 75 distinct columns. This data set was sourced from `https://anlane611.github.io/ids702-fall23/DAA/listings.csv`. The following data cleaning and feature engineering steps where performed (All assumptions have been explained below):

`Filling in missing data`: There was missing data for multiple columns used in this analysis. The missing values in number of bedrooms were filled in using 1. This was done under the assumption that missing entries for bedrooms meant that it was either a hotel room or a loft/studio. 7 entries for the number of bathroom, 239 entries for the host acceptance rate, and 386 super host flags were missing. Unlike the number of bathroom variable, suitable rationales were not discovered to fill in the missing values.

`Amenity score`: The amenity score is a numerical representation assigned to different sets of amenities associated with individual Airbnb listings. These scores are assigned based on their perceived importance, and can be changed dynamically. For instance, amenities like "Wi-Fi" and "Air Conditioning" are considered more crucial and are assigned higher scores, while amenities like "Pets Allowed" are considered less important and receive a lower score. The amenity score is calculated by iterating through each amenity and checking if it is mentioned in the listing's amenities. If an amenity (or amenity sub-string) is found (case insensitive), its associated score is added to the total amenity score for that listing (all other amenities are considered to have zero importance). The amenities that were assigned an importance score are as follows: "wi-fi = 5", "free parking = 4", "air conditioning = 5", "pool = 4", "gym = 3", "kitchen = 4", "Pets Allowed = 1".

`Distance to Downtown`: The distance to downtown variable was calculated using the longitude and latitude columns provided in the data set. The Haversine function was used to arrive at a distance from downtown in miles.

`Number of bathrooms`: The number of bathrooms needed to be converted to numeric from string. This was achieved by iterating through each bathroom string, locating the number sub-string (assuming it was at the start of the full bathroom string) and converting it into a float.

`Price of Rental`: The price of rentals also had to be converted to a float from a string. To do this, the dollar signs and commas where removed from each string, and the resulting string was converted to a float.

`Room type`: The room type variable was converted to a factor variable. This variable is categorical, meaning it represents different categories or groups (e.g., "Private room", "Entire home/apt"). Converting it to a factor explicitly tells R that it should treat this variable as a categorical one.

`Is the host a super host`: This variable is also categorical and was therefore converted into a factor variable.

`Host acceptance rate`: This variable needed to be cleaned from a text string into a number. This was done so that this variable could be used as a continuous variable in fitting the linear model.

Methods

To start, a vanilla linear regression model was fitted using variables mentioned in the methods section of the non-technical report. However, the initial model yielded a relatively low R-squared value, indicating that it did not explain a significant portion of the variability in the rental prices. This prompted further refinement of the model. Below you can find explanations for each category of changes.

`Transformation`: Variable transformation was employed based on the results of a QQ plot, which demonstrated a violation of the normality assumption for the outcome variable. To address this, the outcome variable (price) underwent a logarithmic transformation. This transformation helped align the data with the assumptions of linear regression modeling. All other diagnostic plots were within acceptable limits.

`Interaction terms`: Interaction terms were introduced, particularly for the 'room type' variable. This decision was informed by visualizations, such as scatter plots with fitted lines for each room type. These plots revealed that certain variables exhibited significant interactions with room type, indicating that their effects on price varied depending on the type of room. The variables that benefited from the room type interaction term were number of bedrooms, and number of bathrooms.

`Excluded observations`: All observations associated with shared or hotel room types were excluded from the analysis. This decision was motivated by the sparsity of data within these categories. Sparse data in certain categories can lead to instability in model estimates, potentially resulting in unreliable predictions. Removing these categories aimed to enhance the robustness and predictive accuracy of the model. No observations needed to be dropped based on cooks distance and leverage. All values except one data point where within acceptable bounds. Upon further investigation the data point could not be attributed to faulty data collection, therefore this was not dropped.

`Multicollinerarity`: No adjustments were necessary for multicollinearity. All variables exhibited low correlation coefficients (less than 0.9) and had Variance Inflation Factor (VIF) values below 10. This indicated that multicollinearity was not a concern in this model, allowing for a more reliable interpretation of the relationships between the predictors and the response variable.

The final model was assessed based on the following summary tables.

Table 1: Summary statistics of final linear regression model

Variable	Estimate	Std. Error	t value	Pr(>
(Intercept)	3.7830712	0.0791450	47.799	< 2e-16 ***
# of Bedrooms	0.2225697	0.0126594	17.581	< 2e-16 ***
Room Type (RT)	-0.3039233	0.0866805	-3.506	0.000461
# of Bathrooms (BR)	0.1746483	0.0163816	10.661	< 2e-16 ***
Amenity Score	0.0165973	0.0024247	6.845	9.28e-12
Log of Distance to Downtown	-0.1642382	0.0074918	-21.923	< 2e-16 ***
Is Host a Super Host: False	0.2223081	0.0436765	5.090	3.81e-07
Is Host a Super Host: True	0.3195319	0.0419960	7.609	3.72e-14
Host Acceptance Rate	0.0018811	0.0006539	2.877	0.004048 **
# of BR: Where RT is Private room	0.1072926	0.0634589	1.691	0.090993
# of BR: Where RT is Private room	0.1734080	0.0577896	3.001	0.002717 **

Table 2: More Summary statistics of final linear regression model

Residual standard error	Multiple R-squared	Adjusted R-squared	F-statistic	p-value
0.4071	0.5483	0.5468	353.2	< 2.2e-16

In table 1, each coefficient estimate indicates the magnitude of impact that a one-unit change in the respective predictor has on the log-transformed rental price (this is converted back using the exponent operator). For instance, an increase in the number of bedrooms is associated with a higher rental price.

The 'Adjusted R-squared' value of 0.5468 implies that approximately 54.67% of the variability in rental prices is captured by the included variables. This indicates a good fit for the model, as it explains a substantial portion of the observed price variations. The F-statistic, with a high value of 294.3, reinforces the model's overall statistical significance.

Looking at individual predictor significance, variables such as number of bedrooms, number of bathrooms, total amenity score, and super host flag status exhibit low p-values (<0.05), indicating they are statistically significant in predicting rental prices. On the other hand, the number of bedrooms where room type is Private room, show higher p-values, suggesting this may not be as influential in determining prices.The 95% confidence interval for these coefficients were also observed to be within acceptable ranges.

The residual standard error of 0.4071 indicates that, on average, the model's predictions deviate from the actual rental prices by \$1.52 (exp(0.4071)). This can be considered an acceptable level of error, given the complexity of real-world pricing dynamics.

Conclusions

Overall, this model provides a valuable tool for hosts to estimate rental prices based on a combination of important factors. However, it's important to note that the model's price generation and relationship between predictor and outcome variables are based on the assumption that the relationships observed in the data will hold for future listings. Regular updates and validation against new data will be essential for maintaining accuracy of this model over time.

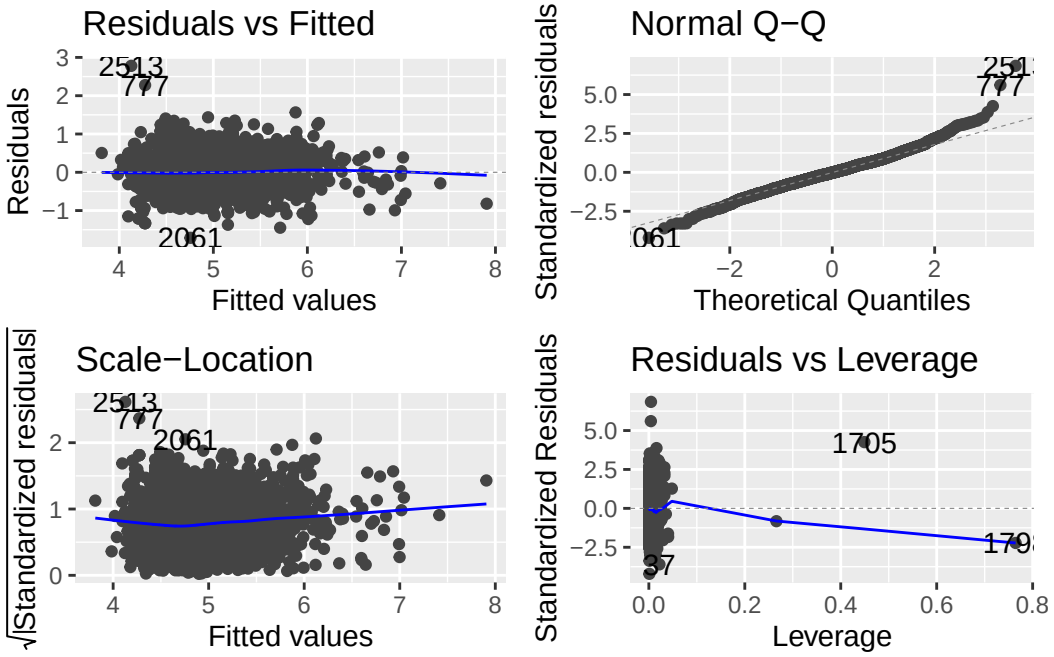

Figure 1: Diagnostic plots for the final model