Airbnb Price Predictor Model

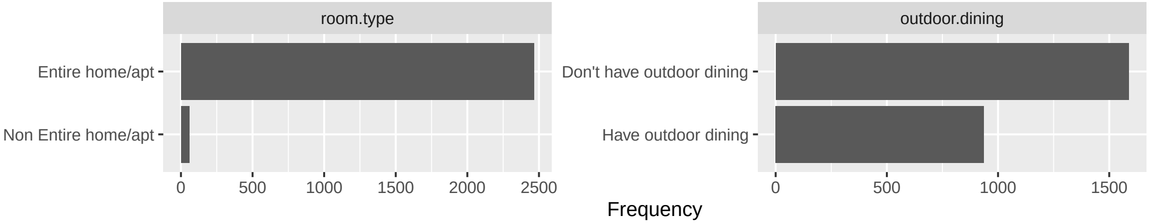
Tina Yi

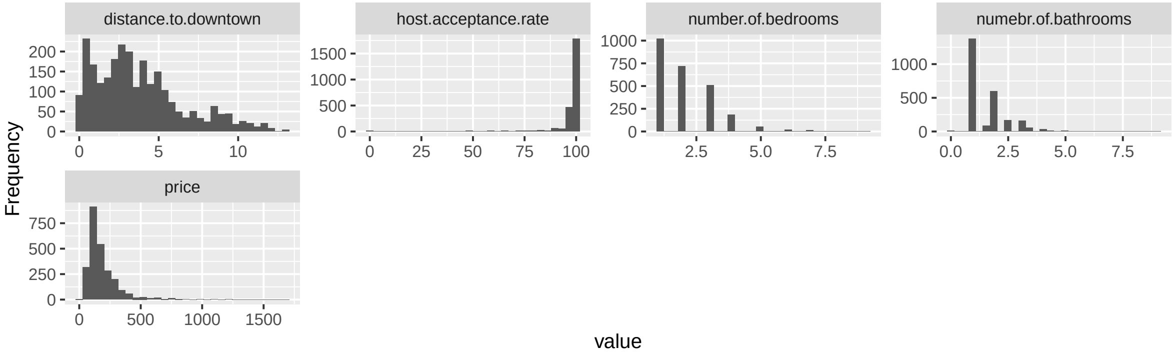
2023-10-01

Report for Airbnb Executives

Introduction: This analysis aims to assist aspiring hosts in establishing pricing for their Airbnb listings in Asheville, NC, by creating a predictive model based on historical Airbnb data. By applying this model, I aim to assist new hosts in understanding and leveraging these variables to set competitive and market-aligned prices for their listings, thereby fulfilling our project goal of aiding hosts in optimizing their pricing strategies.

Method: The goal of the model is to accurately predict the outcome variable price, which represents the daily cost of Airbnb in local currency. To make these predictions, the model considers the following predictor variables, which I believe would have a significant impact on Airbnb prices. (1) **Number of Bedrooms**: Assume that Airbnb with more bedrooms would be priced higher. (2) **Number of Bathrooms**: Assume that Airbnb with more bathrooms would be priced higher. (3) **Distance to Downtown**: Assume that Airbnb with closer distance to downtown would be priced higher. (4) **Host Acceptance Rate**: Assume that Airbnb with higher host acceptance rate would be priced higher. (5) **Room Type**: Assume that Airbnb with entire homes/apartments would be priced higher. (6) **Outdoor Dining**: Assume that Airbnb with outdoor dining would be priced higher.





I established a multiple linear regression model, which investigates the relationship between price and the aforementioned variables, for predicting Airbnb prices. This model provides accurate insights into how various factors above influence Airbnb prices.

Result: I used the Root Mean Square Error (RMSE) value to assess the accuracy of my prediction model. RMSE is a way to understand how close the model's predictions are to the actual results. A lower RMSE means the model's predictions are generally close to the real outcomes, making it a 'good' model. The cross-validation shows that the RMSE value of my model was approximately 0.3954771. This suggests that my model is reasonably accurate in predicting Airbnb prices, with its predictions being relatively close to the actual prices.

Table 1: Model Metrics

Metric	Value
RMSE	0.3954771
R^2	0.5272680
MAE	0.3075177

Here are some examples illustrating the utility of my model:

For an Airbnb listing with 2 bedrooms, 1 bathroom, located 1.5 units away from downtown, categorized as an entire home/apartment with available outdoor dining, and a host acceptance rate of 98%, the model predicts a price of approximately $144.93 per night.

For an Airbnb listing with 3 bedrooms, 2 bathrooms, located 2 units away from downtown, categorized as a non entire home/apartment without outdoor dining, and a host acceptance rate of 100%, is predicted to be approximately $202.66 per night.

Number of Bedrooms	Number of Bathrooms	Distance to Downtown	Room Type	Outdoor Dining Option	Host Acceptance Rate	Predicted Prices
2	1	1.5	Entire home/apt	Have outdoor dining	98	144.9341
3	2	2.0	Non Entire home/apt	Don't have outdoor dining	100	202.6606

Conclusion: The model I created to predict Airbnb prices is working fairly well, but it does need some improvements before it can be used widely.

Currently, I have a lot less information about properties listed as 'Non-Entire home/apt' compared to 'Entire home/apt'. The scarcity of data in 'Non-Entire home/apt' category, having only 61 samples compared to 2466 samples in 'Entire home/apt' category, holds the potential to make our predictions less reliable. Therefore, it is necessary to gather more data samples for 'Non-Entire home/apt' category to get a more accurate model. Also, there are some listings with a '0%' host acceptance rate, and it's unclear what this means—it could be picky hosts, new listings, or maybe just missing information. Understanding and addressing these properly is crucial to avoid any inaccuracies in predictions.

In short, with more balanced and clear data, especially for those special cases, I can make this price predictions more reliable and accurate.

Report for Data Science Team

Introduction: The primary dataset, listing.csv, sourced from Inside Airbnb, is comprised of 3,239 observations, spanning across 75 variables.

```
                                          Descriptions       Value
1                                   Sample size (nrow)        3239
2                              No. of variables (ncol)          75
3                    No. of numeric/interger variables          38
4                              No. of factor variables           0
5                                No. of text variables          23
6                             No. of logical variables           9
7                          No. of identifier variables           2
8                                No. of date variables           5
9             No. of zero variance variables (uniform)           3
10               %. of variables having complete cases    68           (51)
11  %. of variables having >0% and <50% missing cases 25.33           (19)
12 %. of variables having >=50% and <90% missing cases   1.33            (1)
13          %. of variables having >=90% missing cases   5.33            (4)
```

After examining the primary dataset listing.csv, I narrowed the focus to 7 key variables. The outcome variable is price. The predictor variables are room types (categorical), number of bedrooms (numeric) , number of bathrooms (numeric), distance to downtown (continuous), host acceptance rate (numeric) , outdoor dining (categorical) , to develop the Airbnb price prediction model.

Data Cleaning: (1) I began by categorizing room types into four distinct categories: Entire home/apt, Hotel room, Private room, and Shared room. There were no missing values for room type. (2) I eliminated 471 instances where the number of bedrooms was missing. (3) I standardized Shared Half Bath and Half Bath in the bathroom text by replacing them with "0.5 bath". Subsequently, I extracted the numeric values from the bathroom text as the number of bathrooms and omitted one instance due to a missing value. Notably, I treated both private bath and shared bath equivalently. For instance, 1 private bath and 1 shared bath were both counted as 1 bath, as the data samples were limited and the impact difference between private bath and shared bath seemed minimal. (4) I created a new variable to represent the distance to downtown, calculating the distance in meters from a latitude and longitude in downtown Asheville. There were no missing values for the distance to downtown. (5) I refined the price variable by removing the "$" sign, and there were no missing values in price. (6) I created a new outdoor dining variable and populated it with "yes" if "Outdoor dining area" was present in amenities for a given listing and "no" otherwise. I then categorized outdoor dining into two groups: those with outdoor dining and those without. (7) I refined the host acceptance rate variable by removing the "%" sign, finding no missing values in the host acceptance rate.

After the initial cleaning and merging of the data, I conducted a preliminary main regression (price ~ the rest of the variables) for model fitting. I discovered that, after omitting the missing values in bedrooms and bathrooms, only three categories of room types remained (Entire home/apt, Hotel room, Private Room). Due to the disproportionate distribution of these three categories, with 2466 Entire home/apt, 15 Hotel room, and 46 Private room, the standard error for Hotel room and Private room in the initial model fitting was higher than for other variables. Consequently, I decided to combine Hotel room and Private room into a new category "Non entire home/apt." After categorizing this new room type, I obtained two distinct categories: Entire home/apt and Non entire home/apt.

Assumptions: (1) The equivalence assumption between shared and private baths was made due to limited sample sizes, and the potential differences in impact between them were considered minimal. (2) The consolidation of room types was driven by the need to balance the distribution of categories and reduce standard error in model fitting, prioritizing model reliability over the granularity of room type distinctions.

After the final data cleaning and merging, I assembled these variables into a new dataset. This newly structured dataset contains a total of 2,522 observations, comprised of 5 numerical variables and 2 categorical variables, with no missing values present.

Outcome variable:

- *Price: Represents the daily cost of Airbnb in local currency.*

Predictor variables:

- *Number of Bedrooms: Specifies the count of bedrooms available in the Airbnb, ranging between 1 to 9.*
- *Number of Bathrooms: Specifies the count of bathrooms, both private and shared, ranging from 0 to 9 (excluding 8, with decimals like 0.5, 1.5, etc., representing additional bathroom counts).*
- *Distance to Downtown: Measures the proximity of the Airbnb location to the city center.*
- *Host Acceptance Rate: Represents the frequency at which hosts accept booking requests, varying between 0% to 100%.*
- *Room Type: Distinguishes between entire homes/apartments and non-entire homes/apartments*
- *Outdoor Dining: Indicates the availability of outdoor dining facilities, categorized as available or not available.*

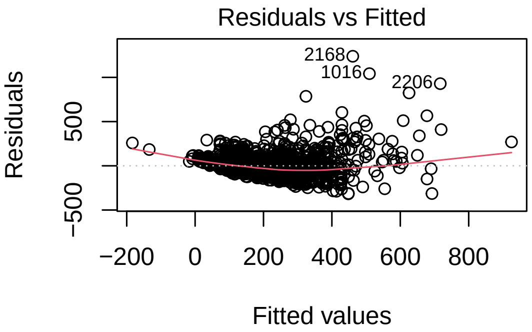
Methods:

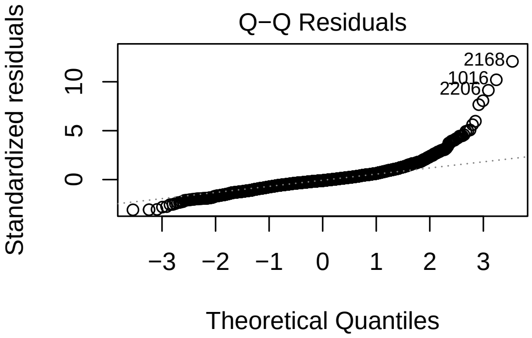
Multicolinearity: The VIF for my preliminary main regression (price ~ the rest of the variables) shows that most of the variables have low multicollinearity, with the adjusted GVIF close to 1, suggesting that they are not correlated with each other. While the number of bedrooms and number of bathrooms variables do show moderate multicollinearity, they are within generally acceptable levels. Therefore, I decide to keep all these variables, since the overall low levels of multicollinearity should not significantly impact the model.

Interaction Terms: In order to determine weather I need to add interaction terms for this model, I created 8 different scatter plots where the outcome variable “price” is plotted against number of bedrooms, number of bathrooms, distance to downtown, host acceptance rate respectively. For the first 4 plots, each plot is colored by room type, showing possible different patterns among different room types. For the last 4 plots, each plot is colored by outdoor dining options, showing possible different patterns among different host acceptance rate.

From these plots, it’s clear that the lines representing different categories aren’t parallel in any of them. This lack of parallelism suggests that the impact of the number of bedrooms, number of bathrooms, distance to downtown, and host acceptance rate on price depends on both the room type and whether outdoor dining options are available. Consequently, it seems logical to consider incorporating interaction terms between each non-categorical variable and both room type and outdoor dining options in the model. However, given that the distance to downtown is the only real continuous variable in this scenario, introducing interaction terms for the number of bedrooms, number of bathrooms, and host acceptance rate may not be meaningful. Therefore, it seems more reasonable to only include interaction terms for the distance to downtown with room type and outdoor dining options.

Variable Transformation: After running my main regression with added interaction terms, I observed the first two diagnostic plots to see if the model violate any of the linear regression assumptions:

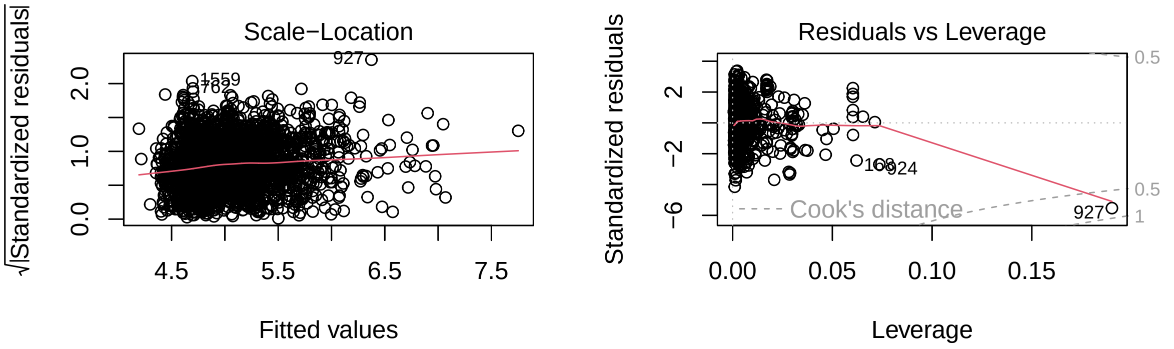




- *Linearity*: From the residual vs fitted plot, it shows the linear relationship between predictor vairables and the mean of outcome variable is violated (non-linear). I need to transform outcome variable (price).
- *Equal variance of errors*: Form the residual vs fitted plot, it shows the equal variance of errors is violated because it demonstrates a cloud shape not equally spread around 0. I need to transform predictor variables.
- *Normality of errors*: From the qq plot, it shows the normality of errors is violated because the tailing points are not clustered around the 45 degree line. After I address the first two assumptions through variable transformation, this assumption should be addressed.
- *Independence of errors*: Observations are independent of each other, which is addressed by the study design.

To address the assumption of linearity, I transformed the outcome variable log(price). I created scatter plots for each non-categorical variable against price to determine if applying the logarithm to any of the non-categorical variables would induce a significant difference in the scatter plots. Observing that the logarithm of the distance to downtown did indeed make a substantial difference, I chose to transform the predictor variable log(distance to downtown) to satisfy the assumption of the equal variance of errors.

Influential Points: After running my main regression with variable transfomrations, I observed the last two diagnostic plots to check if the model have influential points that need to be handled.



From the Residual vs Leverage plot, I observed that point 927 is an influential point as it lies on Cook's distance. I removed point 927 and reran the main regression. After addressing influential points, I began to inspect points with either high leverage or high standardized residuals. I decided to drop point 924, which had high leverage. Upon removing point 924 and observing my regression outcome, I concluded that point 924 is indeed influential as the p-value of 'Non Entire home/apt' becomes more significant (from ** to ***) after its removal, indicating that this point has a substantial impact on the model as a whole. Subsequently, I removed point 168, which also had high leverage, and revisited the model. However, the significance level of the model remained unchanged, suggesting that point 168 is not an influential point as it doesn't have substantial impact on the model as a whole. I decided to keep point 168 and stop my infleuntial points checking at this point. In the end, I removed two influential points, 927 and 924.

Model Assessment: My final regression model has the outcome variable log-transformed listing price, the predictor variables includes the number of bedrooms, the number of bathrooms, the host acceptance rate, the distance to downtown, along with the types of room and outdoor dining availability, which are accounted for as interaction terms with distance to downtown to capture their combined effect on the price. This model utilizes the updated dataset, which excludes the influential points 927 and 924.

		log(price)	
Predictors	Estimates	CI	p
(Intercept)	4.18	4.04 – 4.32	**<0.001**

number of bedrooms	0.22	0.20 – 0.25	**<0.001**
numebr of bathrooms	0.18	0.15 – 0.21	**<0.001**
distance to downtown [log]	-0.08	-0.12 – -0.05	**<0.001**
room type [Non Entire home/apt]	0.19	0.08 – 0.29	**<0.001**
outdoor dining [Don't have outdoor dining]	0.07	0.02 – 0.12	**0.009**
host acceptance rate	0.00	0.00 – 0.00	**0.002**
distance to downtown [log] × room type [Non Entire home/apt]	-0.34	-0.43 – -0.24	**<0.001**
distance to downtown [log] × outdoor dining [Don't have outdoor dining]	-0.11	-0.14 – -0.07	**<0.001**
Observations	2525		
R^2 / R^2 adjusted	0.531 / 0.529		

Given that my model is designed for prediction, I utilized Root Mean Square Error (RMSE) as the metric to evaluate its performance. The cross-validation shows that the RMSE value of my model was approximately 0.3954771. This suggests that my model is reasonably accurate in predicting Airbnb prices, with its predictions being relatively close to the actual prices.

Table 4: Model Metrics

Metric	Value
RMSE	0.3954771
R^2	0.5272680
MAE	0.3075177

*Side Note: In this project, I also developed a secondary regression model incorporating interaction terms for all non-categorical variables with room types and outdoor dining options. This was compared to the main model, which only includes interaction terms for distance to downtown with room types and outdoor dining options. The secondary model yielded an RMSE of approximately 0.3954225, slightly lower than the 0.3954771 of the main model, but the difference is negligible. Given the negligible difference in RMSE, I opted for the main model due to its simplicity and fewer interaction terms.

Conclusion: In assessing the model, it's apparent that it demonstrates satisfactory performance with a commendable RMSE value. However, it also reveals areas requiring refinement before deployment.

One significant concern arises from the disproportionate data samples between the 'Entire home/apt' and 'Non-Entire home/apt' categories. The scarcity of data in 'Non-Entire home/apt' category, having only 61 samples compared to 2466 in 'Entire home/apt' category, holds the potential to skew the model's predictions. Therefore, it is necessary to gather more data samples for 'Non-Entire home/apt' category to get a more balanced data acquisition for an unbiased and robust model.

Additionally, there are ambiguities in the interpretation of a '0%' host acceptance rate. This could imply either a highly selective host, a newly listed property without any inquiries, or possibly, an erroneous or missing entry. A total of 22 samples exhibit this characteristic, and without clarification on the context behind these entries, there's a risk of introducing bias into the model. Finding a way to properly deal with these entries, whether by changing them or removing them, is essential for improving the model's ability to predict accurately and reliably.

In conclusion, while the model shows promise, refinement in these areas, coupled with more comprehensive and clarified data, would be essential to optimize its performance and reliability in price prediction.