Airbnb Price Predictor Model

Jiayi Zhou

2023-10-01

REPORT FOR AIRBNB EXECUTIVES

Introduction

This project creates a model that predicts the prices of Airbnb listings in Asheville, NC. The original dataset is from Inside Airbnb, which initially comprised 3,239 observations and 79 variables. After a data cleaning and organization process, we proceeded to fit the model using a refined dataset containing 2,764 observations and 7 variables.

Methods

Given that our primary objective is prediction, we employed a linear model that incorporates both regular variables including room type, number of bedrooms, number of bathrooms, and the presence of a dishwasher, as well as log-transformed variables including the logarithm of price, distance to downtown, and the minimum number of nights required for a booking. The log transformation ensure that the model aligns with all four statistical assumptions essential for a linear model. The choice of a linear model stems from its ease of use and interpretability, making it well-suited for predicting listing prices.

Results

The model is $\log(\text{price}) = 4.83 + 0.21 \times \text{bedrooms} + 0.16 \times \text{bathrooms} - 0.17 \times \log(\text{distance to downtown}) - 0.33 \times \text{room type} - 0.1 \times \log(\text{minimum nights}) + 0.16 \times \text{dishwasher}$. The model achieved an R-squared value of 0.53 and a root mean square error (RMSE) of 0.40. R-squared measures how effectively the model fits our data, with a value of 1 indicating a perfect fit that explains all variability. RMSE, on the other hand, quantifies the degree to which predictions from the model deviate from actual values. A lower RMSE signifies that predictions closely align with reality, enhancing the accuracy of our predictions. Both R-squared and RMSE values range from 0 to 1. We selected the model with the highest

R-squared and the lowest RMSE among all the models we explored. Based on our chosen model, we predict a listing with two bedrooms, one bathroom, situated 1.5 miles from downtown, requiring at least a one-night stay, and lacking a dishwasher can set its price at $150.

Conclusion

The validity of this analysis relies on the accuracy of the underlying assumptions and the representativeness of the dataset. For future analyses, potential variables for inclusion could encompass the year the listing was constructed, customer ratings, pet-friendliness, and host responsiveness. To further validate the model's performance, it would be valuable to test it using a new, independent dataset. Furthermore, future endeavors should prioritize the acquisition of additional data related to property types beyond entire homes or apartments.

REPORT FOR DATA SCIENCE TEAM

Introduction

General Overview

The dataset is sourced from Inside Airbnb, comprising 3,239 observations and 79 variables. In this project, we focus on seven key variables: the listing's price, room type, number of bedrooms, number of bathrooms, distance to downtown, minimum number of nights for the listing, and whether the place includes a dishwasher. Our objective is to construct a model that accurately predicts the prices of Airbnb listings in Asheville, North Carolina, with the remaining six variables serving as explanatory factors.

Data Cleaning

1. Dishwasher Availability: We extract this information from the amenities variable, categorizing it as a binary variable where 0 signifies no dishwasher and 1 signifies the presence of a dishwasher in the listing.
2. Listing Price: We modify the price variable to eliminate symbols such as '$' and ',' and then convert it into a numeric format.
3. Number of Bathrooms: We address special cases such as 'Shared half-bath,' 'Half-bath,' and empty values by treating shared and non-shared baths as equivalent, assigning both 'Shared half-bath' and 'Half-bath' a value of 0.5, and empty values as missing data. All other cases are converted into numeric values.
4. Distance to Downtown: This is computed using a function provided in the code, resulting in a continuous numeric variable. We assume that the observations are all from the same city, sharing a common downtown location.
5. Room type, number of bedrooms, and minimum number of nights for the listing are categorized as categorical, numeric, and numeric variables, respectively, and are left unaltered.

   We create a new data frame, containing only these seven variables, with missing values omitted.

Methods

Initially, we build a model with no modifications. However, as the table below shown, we observe high standard errors for the 'room type' categories of hotel room and private room, which have only 15 and 65 observations, respectively, compared to 2,687 observations for entire homes or apartments. To address this, we group the 'room type' variable into two categories:

one for entire homes or apartments and another for non-entire places. This adjustment results in a slight decrease in the standard error for non-entire places.

	price		
Predictors	Estimates	CI	p
(Intercept)	39.49	28.68 – 50.30	**<0.001**
room type [Hotel room]	214.35	155.77 – 272.92	**<0.001**
room type [Private room]	162.74	134.03 – 191.45	**<0.001**
bedrooms	38.55	31.85 – 45.25	**<0.001**
bathrooms	69.24	60.73 – 77.75	**<0.001**
dist to dt	-10.84	-12.44 – -9.24	**<0.001**
minimum nights	-1.67	-2.06 – -1.28	**<0.001**
dishwasher	7.87	-2.03 – 17.77	0.119
Observations	2767		
R^2 / R^2 adjusted	0.459 / 0.457		

Subsequently, we create figures where the x-axis represents the explanatory variables, and the y-axis represents the response variable, aiming to identify any potential interaction terms that should be added. Notably, room types exhibit distinct slopes for variables such as distance to downtown and minimum number of nights. However, due to the limited number of observations for hotel rooms and private rooms, we plan to test and address this with cross-validation later. We fit a model with interaction terms between 'distance to downtown' and 'room type,' as well as between 'minimum number of nights' and 'room type.'

Based on the residuals vs. fitted plot, we observe violations of assumptions regarding equal variance and linearity, as the red line follows a curved pattern, and the distribution of points widens as fitted values increase. To address this, we log-transform the response variable 'y,' resulting in improved constant variance in the residual vs. fitted plot.

Next, we generate figures with the x-axis representing the explanatory variables and the y-axis representing the log of the response variable to assess whether any log transformations of the explanatory variables are necessary. We identify that transforming 'minimum nights' is a viable option. Subsequently, we fit a new model and plot the residuals vs. fitted values, revealing improved linearity as the red line flattens and approaches zero.

Then, we proceed to exclude certain observations. The Residuals vs. Leverage plot highlights that observations 98, 2607, and 3113 exhibit high Cook's distance, signifying their influence. Consequently, we remove these observations. The summary table indicates a change in p-values from insignificant to significant for the interaction term 'log minimum nights' and 'room type' with the entire place, leading us to retain the dataset after removing observations 98, 2607, and 3113.

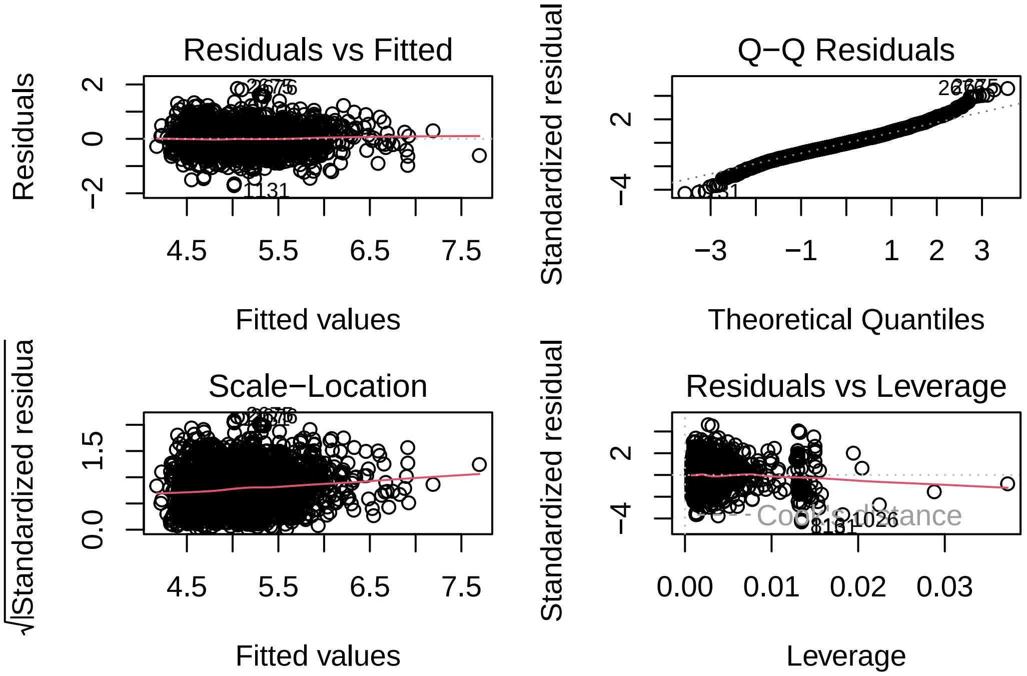
In the context of this prediction task, our objective is to evaluate the model's performance on a test set. Since we lack access to a new dataset, we opt for cross-validation. Given our uncertainty about the inclusion of interaction terms and log transformations, especially due to the limited observations for hotel rooms and private rooms, we employ cross-validation to build models both with and without these modifications.

Based on the evaluation metrics of root mean square error (RMSE) and R-squared provided by all the models, we have selected the following model as our choice: $\log(\text{price}) = 4.83 + 0.21 \times \text{bedrooms} + 0.16 \times \text{bathrooms} - 0.17 \times \log(\text{distance to downtown}) - 0.33 \times \text{room type} - 0.1 \times \log(\text{minimum nights}) + 0.16 \times \text{dishwasher}$. This model exhibits an RMSE of 0.40 and an R-squared value of 0.53. We selected it because it possesses the lowest RMSE and the highest R-squared, and it does not incorporate interaction terms, making it easier to interpret.

The final model summary is presented in the graph below. As all VIF scores are below 5, no adjustments are deemed necessary due to multicollinearity. Furthermore, the assumptions of linearity, equal variance, and normality of errors are satisfied, with no influential data points detected.

	log(price)		
Predictors	Estimates	CI	p
(Intercept)	4.83	4.74 – 4.92	**<0.001**
bedrooms	0.21	0.18 – 0.23	**<0.001**
bathrooms	0.16	0.13 – 0.19	**<0.001**
dist to dt [log]	-0.17	-0.18 – -0.15	**<0.001**
room type combine [yes]	-0.33	-0.43 – -0.24	**<0.001**
minimum nights [log]	-0.10	-0.11 – -0.08	**<0.001**
dishwasher	0.16	0.13 – 0.20	**<0.001**
Observations	2764		
R^2 / R^2 adjusted	0.533 / 0.532		

```
# A tibble: 6 x 2
  Variable             VIF_Value
  <chr>                    <dbl>
1 bedrooms                  3.47
2 bathrooms                 3.18
3 distance to downtown      1.11
4 room type                 1.07
5 minimum nights            1.02
6 dishwasher                1.32
```



Conclusion

In this project, we have created a model to assist in setting prices for Airbnb listings in Asheville, NC. The model utilizes log-transformed variables, including price, distance to downtown, and minimum number of nights for the listing, achieving a root mean square error of 0.40 and an R-squared value of 0.53.

The model incorporates six explanatory variables: price of the listing, room type, number of bedrooms, number of bathrooms, distance to downtown, minimum number of nights for the listing, and the presence of a dishwasher, all of which are used to predict the listing price.

The validity of this analysis hinges upon the accuracy of the assumptions made and the representativeness of the dataset. For future analyses, potential variables that could be considered for inclusion are the year the listing was built, customer ratings, pet-friendliness, and host responsiveness. Furthermore, variables have been adjusted based on a few assumptions, and further refinement may be achieved through additional inquiries and modifications to these variables. Addressing missing values with more detailed information can enhance the model's accuracy.

To further validate the model, testing it with a new, independent test dataset would be valuable. Additionally, future efforts should prioritize gathering additional data related to property types other than entire homes or apartments.