Airbnb Pricing Prediction Data Analysis

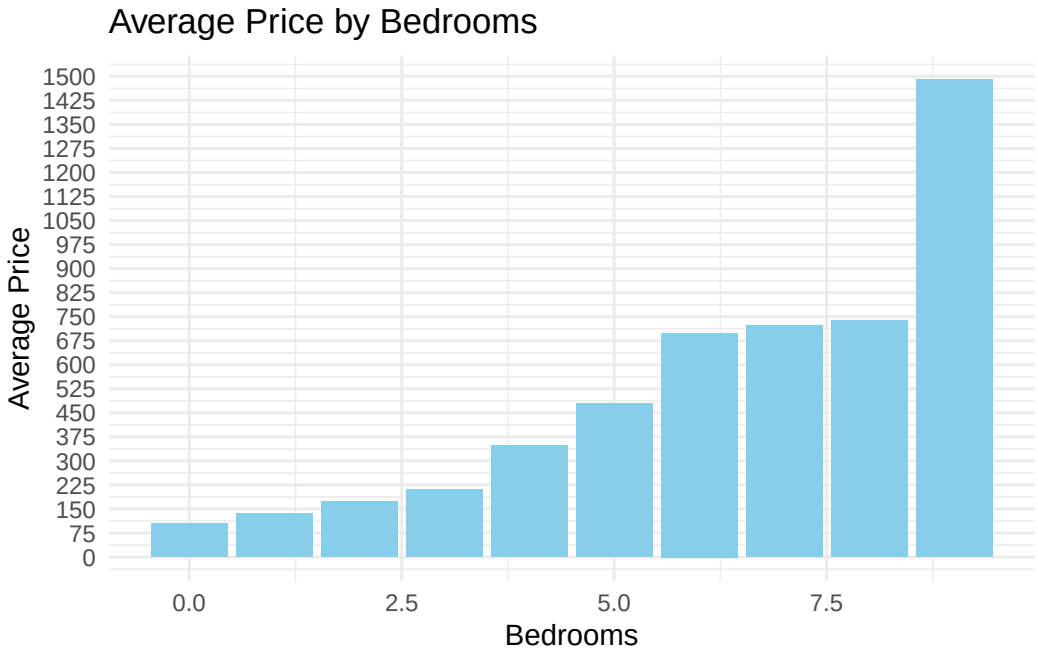
Jiechen Li

Report for Airbnb Executives

Introduction

Airbnb is a platform that allows house and apartment owners to rent their properties to guests for short-term stays. My dataset can be found on Inside Airbnb. The dataset describes the hosts listing activities and metrics in Asheville, NC.

This analysis aims at building a data-driven decision making model for new hosts to set prices based on selected features to maximum hosts income as well as guests needs. The influential features of price in this model are vital for guests looking for a accommodation. In other words, those features will also significant for the hosts to set their price properly.



Methods

Methods of analysis include both exploratory data analysis and predictive modeling. The report illustrates the detailed process of data cleaning, variable selection, and linear model building. The reason why I choose linear model is because it can swiftly adapt to various situations and offers straightforward advice. Therefore, this linear model is beneficial for hosts to predict a proper price so that will be favored by the target guests. The variables are selected from a guests perspective. This means what are the most important features for me when booking accommodation will be evaluated. Therefore, there are nine variables including price, room type, bedrooms, beds, bathrooms, distance to downtown, number of reviews, TV and air conditioner in the linear model.

Results

The key score of the model tells us how well our factors (like bedrooms, bathrooms, etc.) explain the prices. It ranges from 0 (not explaining at all) to 1 (explaining perfectly). Here, the score is 0.5559, so the factors explain about 55.59% of the price differences. That's more than half, which is decent but not perfect.

For example, for each additional bedroom, the hosts might increase the price by around $41, but having a TV only increases it by about $32. Here is the hypothetical listing features:

- *Room Type*: Private room
- *Bedrooms*: 2
- *Beds*: 3
- *Bathrooms*: 2 baths
- *Distance to Downtown*: 5 km
- *Number of Reviews*: 50
- *TV*: Yes
- *Air Conditioning*: Yes

So, for an Airbnb listing with the above features, the projected price would be approximately $1,321.06.

Conclusion

The model explains a bit more than half of the reasons prices vary. This means there might be other factors I haven't considered that also influence the price. In essence, if hosts were listing a room on Airbnb, they probably set a higher price if their property had more bedrooms, was an entire home, was closer to downtown, and had amenities like a TV. But hosts also want to consider other things not in this model to get the best price.

Report for Data Science Team

Introduction

This analysis is centered around the linear regression approach to assist new hosts in price optimization. The predictors in this regression model, which significantly influence the accommodation price, serve as crucial indicators for guests when selecting accommodations. For this analysis, nine pertinent predictors were selected and extracted from the 3,239 observations and 75 variables of the primary dataset, resulting in a refined dataset tailored for my modeling purposes.

Methods

Data Cleaning

Throughout this cleaning process, the aim was to transform textual or unstructured data into a format suitable for statistical modeling. By converting text to numbers, handling missing values, and extracting meaningful features, the dataset was prepared for subsequent regression analysis.

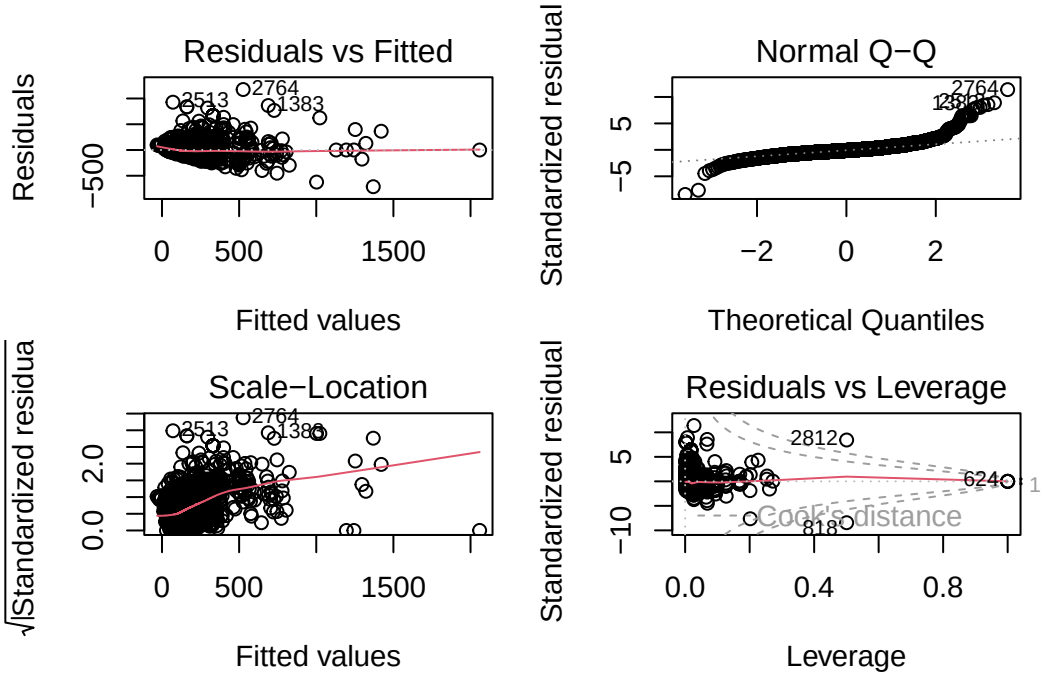
1. **Bathrooms (bathrooms_text)**:
   The variable "bathrooms" are all missing in the original dataset. In this case, I use "bathrooms_text" to represent the number of bathrooms. A new variable "bathrooms_no_NA" was created to extract the numeric part from the "bathrooms_text" variable. Special conditions like "Half-bath" and "shared half-bath" were transformed to the numeric value 0.5. Missing values (NAs) or empty strings in "bathrooms_text" were replaced with zeros. Finally, the cleaned variable was renamed to "bathrooms", replacing the original "bathrooms_text" variable.
2. **Bedrooms (bedrooms)**:
   Missing values in the "bedrooms" variable were replaced with zeros. This assumes that listings without a specified bedroom count can be considered as studio or open spaces.
3. **Beds (beds)**:
   For the "beds" variable, missing values (NAs) were replaced with the value 1, assuming that a listing would have at least one bed.
4. **Room Type (room_type)**:
   The original "room_type" variable had 4 categories. I combined "Hotel room" and "Shared room" because of their little amount, and rename it to "Hotel/Shared".
5. **Price (price)**:
   The "price" variable was cleaned by removing with dollar signs and commas as well as converting it to a numeric format.

6. **Distance to Downtown (dist_to_dt)**:
   A new variable "dist_to_dt" was generated using the "longitude" and "latitude" variables from the original dataset. This variable represents the distance from each listing to a specific downtown location.

7. **Amenities (amenities)**:
   The "Air_conditioning" and "TV" variables are choosen from "amenities" variable. These variables indicate the presence (1) or absence (0) of the respective amenities in the listing. After extracting these features, the original amenities column was dropped from the dataset.

Linear Modeling

A linear regression model "mod_full" was fitted with the response variable as "price" and predictors including room type, bedrooms, beds, bathrooms, distance to downtown, number of reviews, presence of TV, and air conditioning.



1. Residuals vs Fitted: From the plot, there seem to be some patterns, especially a funnel shape, indicating potential heteroscedasticity (non-constant variance of residuals). This can violate the assumptions of linear regression.

2. QQ Plot: From the QQ plot, there's a slight deviation from the reference line, especially at the tails, suggesting potential issues with normality. This could impact the reliability of some statistical tests that assume normality.

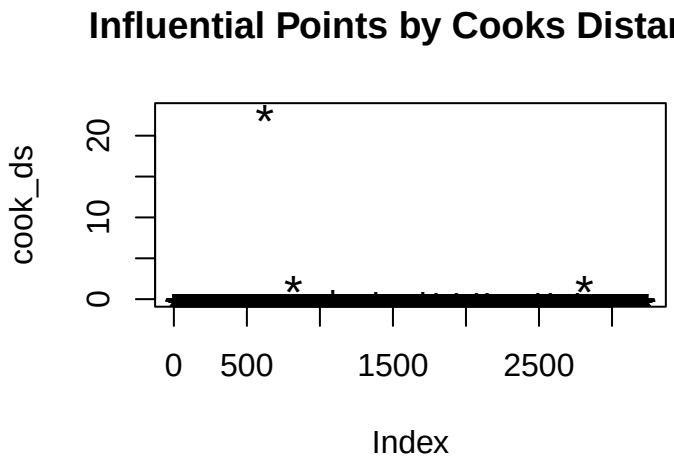
3. Scale-Location: This plot shows if residuals are spread equally along the ranges of predictors.
4. Residuals vs Leverage: The plot shows a few points outside these lines, indicating potential influential observations.

Based on the insights from "mod_full", a refined model "mod_full_2" was constructed excluding beds, number of reviews and air conditioning predictors. However, the performance and assumptions of this model is worse than "mod_full" based on R-squared and Adjusted R-squared. So, I decided to use "mod_full" as the model of this analysis.

In this linear regression model, coefficients are optimized to minimize squared residuals. Residuals are prediction errors; this model typically undervalues by $13.29. The model explains 55.38% of price variability, adjusted to 54.97% considering predictor count. Significant factors include a low p-value (less than 2.2e-16) and an F-statistic of 137.3, denoting predictor significance. Despite the model's fit, 6 values are missing in "bathrooms_no_NA". Still, a few omissions don't notably impact model reliability.

Cross-validation is a technique used to assess the performance of the model by dividing the dataset into multiple subsets. I used 10-fold cross-validation, splitting data into 10 parts, training on 9, and testing on 1. This offers an unbiased performance estimate. Each of the 10 iterations involves almost 3239/10 samples, reflecting the data partition.

Cook's distance is a measure used to identify influential data points in linear regression. Influential points are those observations when they removed, result in a notably different regression model.



Influential points in this plot can unduly influence the regression results, potentially leading to misleading interpretations. Observations that lie above the red line might be considered

influential based on the heuristic threshold set (four times the mean Cook's distance). By removing these observations from the dataset, the model aims to offer a more robust and generalizable representation of the relationships between predictors and the response variable.

VIF is a measure used to detect multicollinearity in regression analyses.

```
                      GVIF Df GVIF^(1/(2*Df))
room_type         7.484827  2        1.654038
bedrooms          3.707203  1        1.925410
beds              1.030222  1        1.014999
bathrooms        22.540706 24        1.067055
dist_to_dt        1.087234  1        1.042705
number_of_reviews 1.107478  1        1.052368
TV                1.180461  1        1.086490
Air_conditioning  1.032931  1        1.016332
```

1. Room Type (room_type): Given that the adjusted GVIF is 1.65 and the predictor has 2 degrees of freedom, it suggests some correlation with other predictors, but it may not be severe.
2. Bedrooms (bedrooms): With an adjusted GVIF of 1.93, there's an indication of some multicollinearity.
3. Bathrooms (bathrooms): Despite a high GVIF, the adjusted GVIF is around 1.07, which suggests that multicollinearity might not be severe when considering the categorical nature of this predictor.
4. Others: Predictors like "beds", "dist_to_dt", "number_of_reviews", "TV", and "Air_conditioning" have adjusted GVIF values close to 1, suggesting little to no multicollinearity.

Conclusion

The model diagnostics, such as R-squared and Adjusted R-squared, indicate that the model explains approximately 55.38% of the variability in the prices, which suggests a moderate fit to the data. The p-value associated with the F-statistic is close to zero, further confirming the overall significance of the predictors.

In conclusion, while the regression model provides valuable insights into factors influencing Airbnb prices, careful consideration of its limitations and assumptions is essential. It's advisable to further refine the model, perhaps by addressing multicollinearity, imputing missing values, or incorporating other relevant predictors, to enhance its predictive accuracy and reliability.