Individual Assignment 1: Airbnb Pricing

Kelly Tong

Report for Airbnb Executives

Introduction: Dataset Description

This report aims to support Airbnb in predicting prices for Airbnb listings in Asheville via modeling the available data. Discussion will also include effects of various Airbnb features on prices setting decisions since understanding the correlation would benefit building the prediction model. The data is sourced from "listings.csv." This dataset provides a comprehensive view of Airbnb listings, caputuring various aspects like pricing, reviews, number of bedrooms, number of bathrooms, amenities and other features. Specifically, there are 2777 listings with unique IDs in total.

Methods

Linear regression is used for its straightforwardness and clarity in highlighting variable relationships. In predicting Airbnb prices, it assumes a direct link between features like location, bedrooms, and amenities to the price. Each feature gets a coefficient, indicating its impact on price. For example, the coefficient for 'bedrooms' shows the price rise for an extra bedroom. By combining all feature contributions, the model estimates Airbnb prices, support understanding feature significance and predict prices for new listings effectively.

Variables included for prediction via the linear regression model include: room types, bedrooms, bathrooms, review score value, pets allowed, accommodates, and distance to downtown. Room type can hint at the listing's size and, consequently, its price - an entire house typically costs more than a shared room. The number of bedrooms and bathrooms usually indicates the listing's size and capacity, influencing its price. Review scores reflect the listing's quality and guest satisfaction, directly impacting its price. The pet-friendly feature might affect pricing based on room type. The accommodation capacity denotes the number of guests a listing can house - a larger capacity often means a higher price. Lastly, proximity to downtown is crucial for many travelers due to accessibility to landmarks and conveniences, making it a significant price determinant.

Results

The primary metric used to assess the model's performance is R squared, which represents the proportion of the variation in log of price that's explained by the selected feature variables. The R squared value is 0.5346, meaning the model explains approximately 53.46% of the variation in Airbnb prices. This is a decent value, suggesting the model has captured over half of the variability in prices using the chosen features. Residual standard error (RSE) is another important metric that measures prediction deviation from the actual values. It is 0.4173, which, given the scale and nature of the log-transformed data, suggests that our predictions are reasonably close to the true log_prices, though there is still some error.

Table 1: Example Prediction of Price using the Prediction Model Output

Predictor Variable (selected features of Airbnb)	**Estimate Coefficient (model output)**	**Example of using model output to predict price**	**Computational Result(example value x coefficient)**
Bedrooms	0.167431	2 bedrooms	2 x 0.167431
Bathrooms	0.023219	2 bathrooms	2 x 0.023219
Room Types (Other room types)	-0.168288	Other room type	-0.168288
Review Score Value	-0.011898	4.6	4.6 x (-0.011898)
Accommodates	0.084580	5	5 x 0.084580
Distance to Downtown (in miles)	0.083075	3.5miles	3.5 x 0.083075
Distance to Downtown (log)	-0.617466	log(3.5)	log(3.5)x(-0.617466)
Distance to Downtown and Room Type combined (interaction term)	-0.009825	3.5*Other room type	-0.009825 x 3.5
Pets allowed	-0.016944	pets allowed	-0.016944
Pets allowed and Room Type combined (interaction term)	0.115209	pets allowed*Other room type	0.115209
Result (sum of computational result + intercept) (log price)			5.440916
Predicted Price (Dollar)			230.653365

An example of how the model can be used to predict the price is provided. *Table 1* summarized the estimated coefficient for each predictor variable as well as the processed computation for predicting price. First of all, estimated Coefficients are multiplied to example input for numeric input. They are then sum up and added with constant values (output for categorical variables and the intercept). After that, to compute the predicted price in dollars, we need to take the exponential of log predicted price (5.440916). The result is 230.653365. Consequently, for this example, the final predicted price is approximately 231 dollars.

Conclusion

The model does provide valuable prediction of price, which can be implemented in reality to some degree. However, it can be improved by expanding the sample size and collecting additional data. Possible additional data collection include cancellation policies, local event data, and security features. Cancellation policies and security are additional features that clients might care about. Local event data can imply whether it would become a popular place for travel and thus lead to increase in price.

Report for Data Science Team

Data Description

This report aims to explain the insights of model assessment and building process to a data science team. The data is sourced from the given dataset "listings.csv." This dataset provides a comprehensive view of Airbnb listings, caputuring various aspects like pricing, reviews, number of bedrooms, number of bathrooms, amenities and other features. The data spans a wide range of property types, prices and locations, which reflects the diversity of Airbnb offerings. Specifically, there are 2777 listings with unique IDs in total. The prediction model will include both categorical and numeric variables, some of which need to be created and defined mannually. Variables included are price, room type, number of bedrooms, number of bathrooms, accomodates, review score value, pets allowed, and distance to downtown. Cleaning, definitions and selection of these variables will be explained below.

Price: While price needs to be a numeric continuous variable for mathematical operations, it is stored as a character variable in the source dataset. It can be altered into a numeric variable in two steps. Firstly, extract the numeric information from the original character vector and stores it into a new variable (price 1 for example.) Then, convert the variable to numeric mode using the as numeric function. After replacing the original price variable with the cleaned one, we can predict the price later.

Room Types: Room types included in the original dataset involves "Entire home/apt," "Shared room," "Private room, and"Hotel room." Since some of these room types only have a few observations, we will re-categorize and merge room types into "Entire home" and "Other room types." This will simplify the model for easier interpretation. It also improves generalization, allowing the model to be applied to new data. More importantly, it avoids over-fitting since the model is less likely to over-fit with fewer categories.

Bedroom number: Bedroom number variable involves 471 null value. These null values are replaced with 1 bedroom. This is because the null value in bedroom number is mostly caused by characteristics of hotel rooms and private rooms in studio style. While these types of room do have a bedroom (place for sleeping), they are different from the typical bedrooms. This might be the reason for null values in the dataset. Hence, we assume bedroom number for all the null value to be 1. This is an assumption developed for replacing null values with meaningful numeric values, which could be used in the prediction process.

Bathroom number: The number of bathroom variable in the source data is empty. Information on number of bathroom can be extracted from the bathroom text variable. Specifically, the information for 0.5 is stored as half bathroom/shared half bathroom in the text. Hence, it needs to be dealt with separately. This can be processed through the string substring and as numeric function.

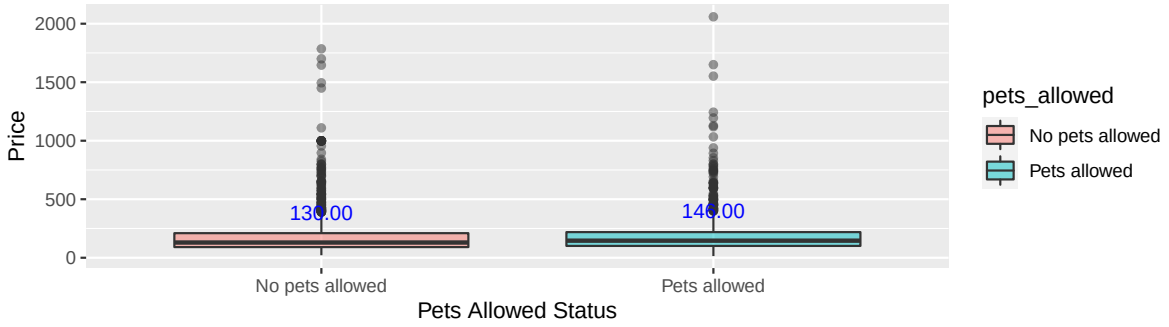
Review Score Value: Review Score Value is a variable that can directly reflects past consumer satisfactory with the room. It will generally influence future clients' expectations and

selection decisions a lot. This variable is also proven to be highly influential to price via Recursive Feature Selection (RFE).

Accommodates: This variable provides information on the number of people that the room can contain (as suggested by household and Airbnb.) Rooms that can accommodate more tend to be larger and thus more expensive. This variable is also suggested by Recursive Feature Selection as one of the influential factors.

Amenity Features: Pet Allowed: Pets allowed is a new variable that is manually created and defined. It stores information on whether the Airbnb allows pets. This information can be found in the amenities variable in the data. I defined an Airbnb as pets allowed when text such as "pets allowed," "pets friendly," "dogs," and "cats" can be searched in the amenities content. As a result, there are 1004 Airbnbs which allow pets and 1984 Airbnbs that disallow pets. From Figure 1, we can see that the median prices for pets allowed (median price equals 146) Airbnb are generally higher than those that do not allow pets (median price equals 130).

Figure 1: Relationship between Pets Allowed and Price



Distance to Downtown: The distance to downtown variable stores information on the computed distance in miles between each Airbnb listings and downtown (with longitude -82.55481168521978 and latitude 35.59701329976918). This variable is manually created and defined.

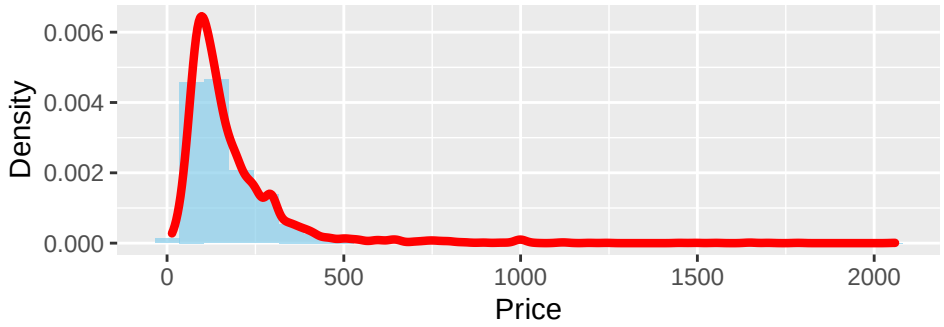
Variable Selection: RFE is used to select the top variables correlated with price. In this case, variables with high collinearity are excluded. For example, review scores rating is generally linear with review scores value. Hence review scores rating is excluded from the RFE test and only review scores value is kept. The resulted top 5 relevant variables are bedrooms, bathrooms, distance to downtown, review score values, and accommodates. These are all included in the prediction model.

Methods and Modeling

Variable Transformation: The price variable is transformed into log price for model prediction due to skewness in the distribution of price *(Figure2)*. This skewness to the right could lead to a few challenges and implications for statistical modeling and interpretation. These

include misrepresentation of central tendency and violation of the linear regression model assumptions. The mean, in a right-skewed distribution, will be influenced by extreme values. Hence, it does not provide an accurate representation of the data. Moreover, this could potentially violate assumptions for statistical tests. For example, linear regression model assumes that data is normally distributed. This is violated by a right-skewed distribution. Hence, we need to transform price into log price to avoid these problems.

Figure 2: Distribution of Price



Distribution of variable distance to downtown also demonstrates right skewness. The same problems, as for right-skewed price distribution, are caused by a right-skewed distance to downtown variable. Therefore, it is also transformed into log value. It has to be noticed that all distance to downtown values are increased by 1 before taking the log. This is because some values are less than 1 mile, and this will lead to negative log results. Negative distance does not match the context since distance is a magnitude variable here. Adding 1 to each value will successfully avoid negative log values and will not impact the relative distance to downtown for each Airbnb.

Interaction Term-Distance to Downtown and Room Type: This interaction term considers that the effects of distance to downtown could vary depending on the types of room. As the Airbnb is more closely located to downtown, the supply of entire homes or apartments could decrease and thus command a higher price. In the meanwhile, other room types, which include shared rooms and private rooms, will not reflect as much of a price increase based on their proximity to downtown.

Interaction Term-Pets Allowed and Room Type: This considers that the cost impact of allowing pets might differ by room type. For example, allowing pets in shared spaces might affect price differently than in entire rooms. This term helps assess how room type modulates the price effect of allowing pets.

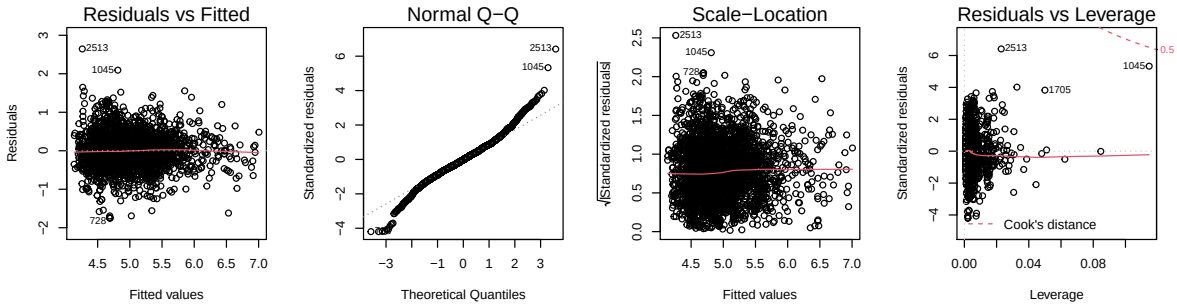
A linear regression model is used to predict Airbnb prices based on the selected variables: bedrooms, bathrooms, room types, review score value, distance to downtown, pets allowed, and accommodates. This offers insights into how these factors influence price. Interaction terms in the model highlight nuanced relationships between variables. The model's summary indicates

the significance and impact of each feature. To ensure robustness and avoid overfitting, cross-validation is employed.

Validity and Conclusion

The regression model predicts Airbnb prices using various property attributes and accounts for about 53.46% of the price's variability, as indicated by multiple R-squared. Notably, properties with more bedrooms, bathrooms, or higher accommodation capacity tend to be pricier. Distance to downtown also impacts the price: the farther away, the lower the price. Interestingly, "Other Room Types" typically have lower prices than entire homes. The effect of allowing pets on the price varies depending on the room type. Surprisingly, higher review scores correlate with lower prices, though this relationship isn't statistically significant (high p-value). This unexpected finding might result from the limited range in review scores present in the dataset. Most score values tend to fall between 4.5 to 4.8 for the data. This leads to lack of variability which could restrain the model from interpreting linear relationship between review score value and log of price.

We can also look at the residual standard error and F-statistics for evaluating model fitting. The residual standard error is 0.4173 which measures the average difference between the observed and predicted values. *(plots for residuals included below)* The p-value for F-statistics is very low, indicating that the model is statistically significant overall. Cross Validation is done with 10 folds. The result present a RMSE of approximately 0.418 and R squared of 0.53. The standard error maintain almost the same in comparison with the linear regression model without cross validation. This implies that the model avoid overfitting relatively well.



Last but not least, there are a few limitations a linear regression model. Linear regression assumes a straight-line relationship between predictors and the outcome, which might not always reflect complex systems like pricing. The model can be skewed by outliers. For better predictions, we could increase the sample size to mitigate bias and outlier effects. Despite these limitations, the model offers valuable insights into Airbnb pricing factors.