Project 2: Predicting the Price of an Airbnb Rental

Scott Kroeger

Project Design

I used data gathered from the Airbnb website to predict the nightly price of the rental property. This data could be used by property owners that are considering offering their properties up for rental on the Airbnb website. If this data was aggregated for each city, potential travelers looking to determine where to travel to based on the price of housing in the area.

My goal was to collect the data from the Airbnb website for each rental that appeared in the Bogota area rental search, and to use the combined features to predict the price using a linear regression model.

Modeling was performed using the Scitkit-learn Linear Regression model. Additionally Lasso and Ridge Regression modeling was also performed as a secondary measure to see if performance could be improved using those models. A range of alpha values for tried for both Lasso and Ridge Regression models.

Using the Linear Regression model I was able to determine that the most important features in determining price of a rental were the number of rooms, followed by the number of bathrooms, free parking, and location.

Tools

- Python
  - Data Storage: csv files, pickled objects
  - Data Analysis: scikit-learn, pandas, numpy, Jupyter Notebook
  - Presentation: matplotlib, Powerpoint

Data

The data consisted of information scraped from the Airbnb website. The data was gather by loading Airbnb webpages within defined price ranges, dates and location. After the data was gathered it was converted to a pandas DataFrame object and saved to a comma separated values file.

The features used are provided in the Appendix. I used all features listed in the provided table. The features were extracted from a JSON object that was part of the webpage, and did not require a much reformatting. The distance formula used to calculated the linear distance between the property and the historic city center was the Equirectangular approximation found at the https://www.movable-type.co.uk/scripts/latlong.html website.

What I Would Do Differently Next Time

I would have spent more time scraping data from Airbnb vs. attempting to improve the outcome of the predictions via tuning the model. I was also reluctant to add more categorical data because I was afraid it would make my model worse. Next time I would gather as much data as possible, and then add or subtract categorical variables if necessary. Not wanting to use categorical data also deterred me from trying to add in categorical variables for locations which I

think would have had a positive impact on the model. I would also go through my process step by step with a peer or instructor to get any further feedback that they may be able to provide in order to improve my process or point out any steps I might have missed like taking the log of my output variable.

Appendix I: Data

Variable	Type	Description	Used for model
bathrooms	Float	The number of persons that can stay at the rental	Y
bedrooms	Int	The number of bedrooms	Y
Beds	Int	The number of beds available	Y
cleaning_fee	Float	The cleaning fee charged by the owner	Y
is_superhost	Categorical	Whether or not the owner is classified as a super host by Airbnb	Y
person_capacity	Int	The number of guests that can stay in the rental at one time	Y
picture_count	Int	The number of images the rental has on the Airbnb website	Y
preview_amenity_names_Free parking	Categorical	Has free parking or not	Y
preview_amenity_names_Kitchen	Categorical	Has a kitchen or not	Y
preview_amenity_names_Washer	Categorical	Has a cloths washer or not	Y
preview_amenity_names_Wifi	Categorical	Has WIFI or not	Y
linear_distance	Float	Distance to the historic city center	Y
price_numeric	Float	The nightly cost of the rental (target variable)	N