Project Overview and Analysis Write-Up

Data

The dataset we are working with comprises Airbnb listings from various New York City neighborhoods spanning 2008 to 2019. It includes feature variables such as precise geographical locations (longitude and latitude), neighborhood, city, unique identifying IDs, listing name, host name, host ID, number of reviews, room type, and availability. These features serve as independent variables in our analysis. Our main objective in this project is to construct a regression model for predicting price, the target variable, which represents the listing's rental price. Through this analysis, we aim to uncover the factors influencing Airbnb listing prices in different neighborhoods of New York City during the specified time period.

Three distinct Airbnb listings are displayed below. The first listing, with the unique ID a710172032669867487, is managed by Kyle (ID: 154834539) and is located in Long Island City, Queens. This listing offers an entire unit for rent with a minimum 2-night stay requirement, and it is available for just 69 more days. On average, it receives 5.42 reviews per month, and the nightly price is set at $199. The second listing, a49649623, is an entire home available for a minimum 6-night stay in Hell's Kitchen, Manhattan. This listing has accumulated a total of 28 reviews, and the listed price is $117 per night. Lastly, the third listing offers a private room in Brooklyn, managed by Lisa, with a unique ID of a3207986. The nightly price for this listing is $125.

For creating a regression model to estimate the price of Airbnb listings, we identified three columns that we hypothesized will be useful: room type, neighborhood group, and minimum nights stay required. The room type column is expected to be closely correlated with the listed price because the type of room available for rent influences pricing. Different room types, such as entire homes, private rooms, or shared rooms, are often associated with different price values. Including this column will allow us to understand pricing variations based on room types. The neighborhood group column is important since it provides us with the geographical location of the listing and plays a pivotal role in determining the price. Different neighborhoods in New York exhibit variations in housing availability and local amenities, which can impact pricing of listings in different ways. This column is considered important to understand the changes in prices based on location or neighborhood groups in New York. The minimum nights stay required column is an important predictor in this scenario since it is possible that higher minimum night stay values may have an impact on the prices of the listing.

Methods

To conduct data preprocessing by performing one hot encoding on two categorical variables, neighborhood_group and room_type. Since we felt that these variables were useful in modeling price of listings, we continued with this step to convert them into numerical values. By doing so, we created a binary column for each category within the neighborhood group and room type. We created a function cap_outliers to remove our outliers for each category. Values above Q3 + 1.5 * IQR and below Q1 - 1.5 * IQR were considered outliers and were replaced with their upper or lower limits. We also allocated 80% of the data to the training set (X_train and y_train) and the remaining 20% to the test set (X_test and y_test). Next, we selected a group of features that we felt were most important in modeling the data. Our features chosen kept changing as we logically thought through whether or not a column will affect the price. After conducting multiple tests, we concluded that the most influential features included: capped_minimum_nights,

capped_calculated_host_listings_count, Bronx, Manhattan, Queens, Staten Island, Entire home/apt, Private room, Shared room, Hotel room, latitude, and longitude as our selected parameters.
We selected a linear regression model to understand the linear relationship between the defined features and price of the listing. The model was trained on our training dataset as provided. To assess the effectiveness of our model, we used the mean squared error as our evaluation metric. This is because it is a standard in linear regression. MSE is the root of squared difference, which means that we are able to get the absolute difference, regardless of direction. The lower the value for the MSE, the more precise the model's predictions are. This metric allowed us to see how well our model's predictions aligned with the actual prices in the dataset. In this case, we did not perform hyperparameter tuning but focused on experimenting with different features to identify those that best predict the outcome. Utilizing the training data, we established a regression line that allowed us to make predictions for the validation set.

Results

As mentioned previously, we did not try different models but rather tested different features within the same linear regression model. Our previous scores were all relatively consistent and scores ranged from 123 to 125. In the future, it may be worthwhile to experiment with different regression models to understand which will best fit the data at hand.