Predicting Airbnb Listing Price | Exploratory Data Analysis

Claire Miles

After cleaning the data and starting to craft the beginning of my data story, it was time to enrich my exploration with further analysis. My exploratory data analysis included the following steps:

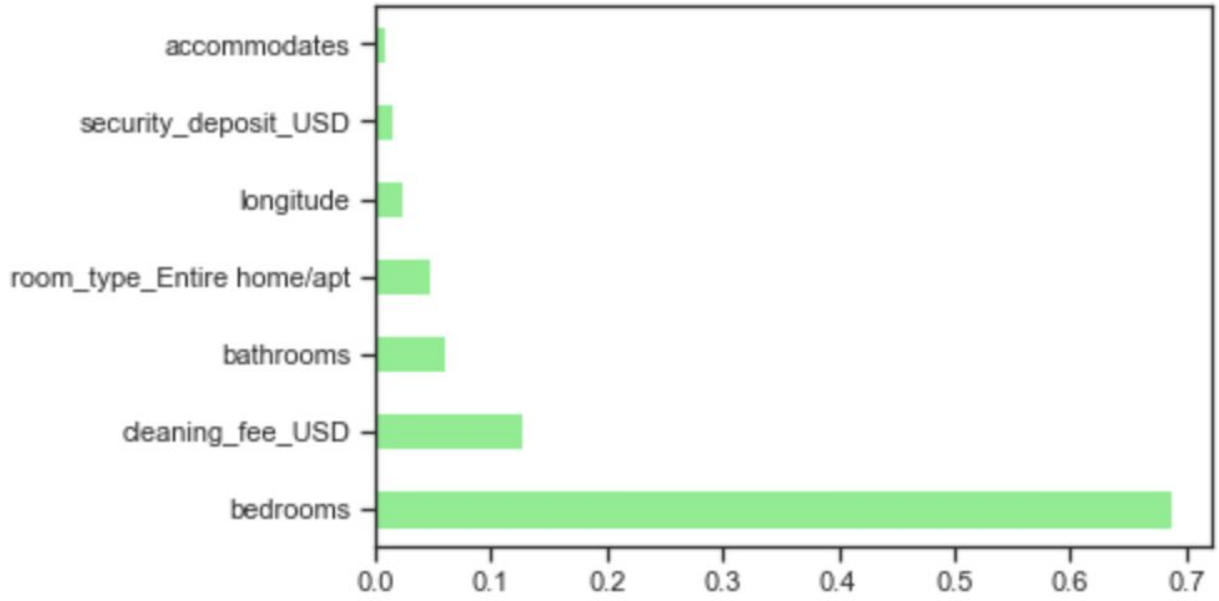
1. Identify important features with a first-pass random forests model.
2. Look at the distributions of the most important features.
3. Re-evaluate the original project question based on the results.

Once again, I focused on the listings dataset in this notebook, as the listing data hosts all numerical features that the dataset currently has. The reviews data, which contains the review text for different listings, will be interpreted using NLP later on in the project process, as will other text data in the listings file.

Random Forests:

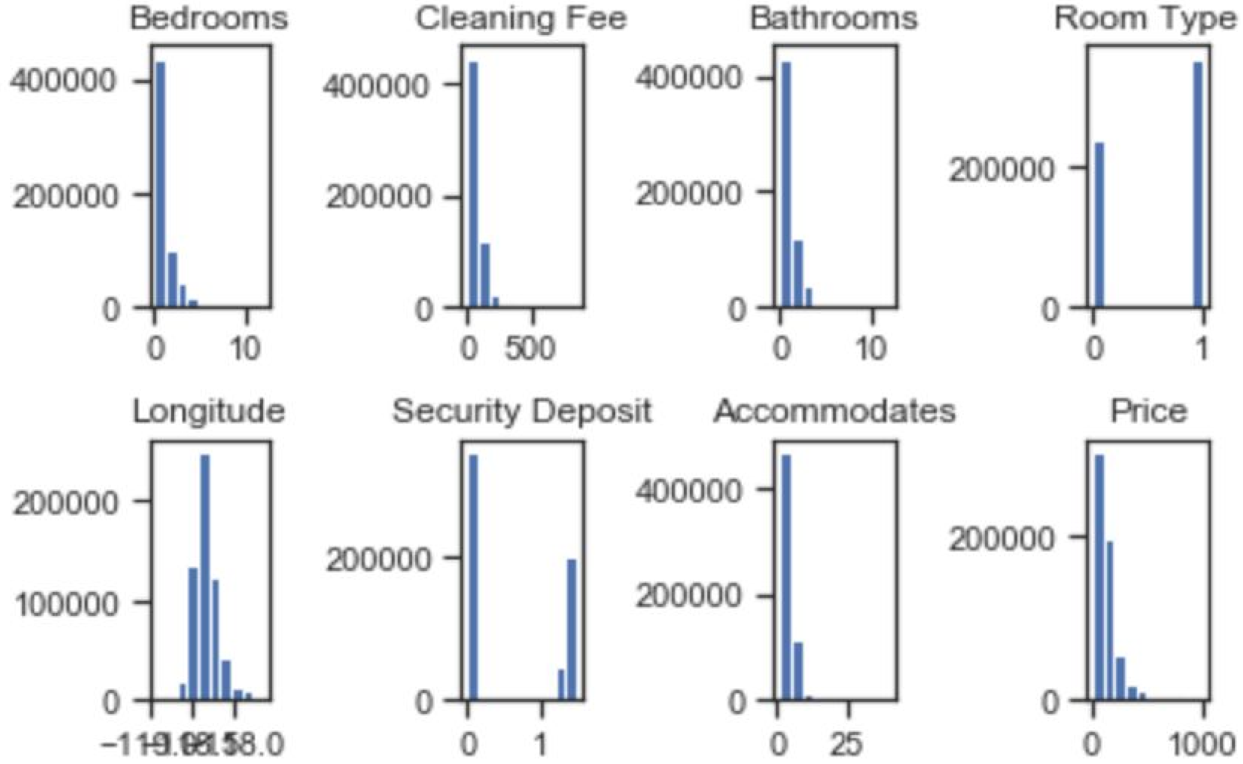
I briefly prepared the data for the random forests training by converting certain attributes of the dataset to categorical data type, turning booleans into 0/1 values, and selecting only the numeric, datetime, and categorical data types in the dataframe. To get rid of NaN values, I used pandas interpolate function to fill in the values linearly (I'll give more thought to interpolation in the training of the final model, but here I'm just looking for a crude first pass). Then, I trained a RandomForestRegressor classifier on one-third of the data and plotted the features by their importances.

Random Forest Feature Importances



The random forest found the most important features to be bedrooms, cleaning fee, bathrooms, room type-entire home/apartment, longitude, security deposit, and accommodates. It's important to note, however, that this was a very simplistic first-pass at the data, and that the feature importance algorithm for random forests shows some bias towards continuous features or high-cardinality categorical variables.

Next, I plotted histograms of the features and the target variable to get a better picture of their distributions:



It appears that longitude is the only normally distributed feature, with bedrooms, bathrooms, cleaning fee, accommodates, and price being right skewed. This makes sense since Airbnb listings are spread throughout the entire city, with some cities being a bit more popular than others. Also, most Airbnb listings are small, affordable, and for hosting just a few people - the smaller amount of larger listings creates the skew of the histogram. Room type is a categorical variable that has been transformed in the one hot encoding process, therefore the histogram just shows that there are more entire homes/apartments than not. For security deposit, it looks like most listings do not have one, but those that do may give us important insights into a listing's price.

Re-evaluation of project question:

The biggest finding of my statistical analysis is that I've identified several of the most important features in the dataset and looked at their distributions. This will inform which features to pay the most attention to in the next stage of the project.

Moving forward, the machine learning analysis will likely offer more complex insights than this statistical analysis, especially as I incorporate information from the reviews data. It may be the case that some features that are not significant in this analysis play a larger role down the line.

This analysis does not alter the initial project question of trying to decipher a listing's price from its other features.