Exploratory Analysis of San Francisco Airbnb Data and Prediction of Optimal Price for a Property

Aim: I have analyzed over 7,834 listings in the San Francisco area in order to better understand how the use of listing attributes such as bedrooms, location, ratings, and more can be used to accurately predict the optimal listing price for both the host and guest. Holiday and seasonality is another useful component that can attract more customers and drive higher prices, but it is unclear how much of a premium one should pay per holiday. With better price suggestion estimates, Airbnb home providers can reach an equilibrium price that optimizes profit and affordability. The objective of this project is to build a model that predicts the optimal price of a property taking into account listing features and seasonality. The end goal is so users can understand what features of an Airbnb listing are most important as well as how prices should be fluctuating based on seasonality.

The Project has three sections of Jupyter notebook code files:

1. Data Cleansing:

- Data Cleaning Listings Code File:

Listings.csv contains detailed listings data, including various attributes (features) of each listing such as location, number of bedrooms, bathrooms, type of bed, reviews, etc. This code file cleans the data by handling Missing values, skewness of the data, z- score normalization of data etc.

- Seasonality Cleaning Code File:

This file is to clean the Calender.csv file, which has listings, availability and their price.

- Zip code Clustering Code File:

This file is to take zip codes and cluster them into neighborhoods. This is better than using the original neighborhoods column as there were many missing values and non-uniform neighborhood names. By converting by cleaned zip code, we can ensure that the clustering by neighborhoods is more accurate.

2. Data Exploration

This section is to explore the data to get a better understanding of the relationship between the observations as well as the relationship between each observation and predictor.

- Data Exploration Listings Code File:

- Check for collinearity among the features
- Visualize relationship between each predictor and price
- Visualize the target variable to identify any skewness and any necessary transformations
- Visualize the supply of Airbnb homes by location

- **Data Exploration Listings Code File:**

The goal of the seasonality analysis is to flush out the model that we used for predicting AirBnB pricing. Prices obviously change over time, so this added level of specificity makes the model that much more usable. This analysis is to discover any seasonal trends in the year for pricing analysis.

2. Data Modeling

- **Data Modeling Baseline Code File:**

- I have implemented Various Machine Learning Models to predict an optimal price for an Airbnb and analyzed their performance. Below are the algorithms I have implemented for prediction of price:

- Linear Regression Analysis with Untransformed Response
- Linear Regression with Log-Transformed Response (Predictor Variable is log transformed)
- Ridge Regression with Untransformed Response
- Ridge Regression with Log-Transformed Response
- Lasso Regression with Untransformed Response
- Lasso Regression with Log-Transformed Response
- Polynomial RidgeCV Model
- Random Forest Regressor
- Random Forest Regressor Untransformed Response
- Random Forest Regressor Log-Transformed Response

A**nalyze models by their Median Absolute Error.**

- Median Absolute Error Untransformed Response
- Median Absolute Error Log-Transformed Response

We now try a regression solely on single listings, as this is where the majority of our listing data lies; namely, listings that only had a bed of 1.

- Ridge Regression Single Listing Log-Transformed Response
- Lasso Regression Single Listing Untransformed Response
- Lasso Regression Single Listing Log-Transformed Response

A**nalyze models by their Median Absolute Error.**

- Median Absolute Error Single House Listings Untransformed Response
- Median Absolute Error Single House Listings Log-Transformed Response

Data Modeling Baseline Cluster Code File:
Assessed the results of the zip code -> neighborhood clustering conversion, running the same models earlier. Data Modeling: Airbnb Listings and comparing the results.

- Ridge Regression with Untransformed Response
- Ridge Regression with Log-Transformed Response
- Lasso Regression with Untransformed Response
- Lasso Regression with Untransformed Response

A**nalyze models by their Median Absolute Error.**

- Median Absolute Error Single House Listings Untransformed Response
- Median Absolute Error Single House Listings Log-Transformed Response

Seasonality Modeling:

The general idea behind this analysis is as follows: Prices are aggregated by weekday for each listing. Then, we normalize each listing's price by the Monday price to find an average multiplier for each listing for each day. Then, for each day, we average across all listings to get a final average multiplier for each day. Lastly, we compare these predictions to a subset of the listings.