Machine Learning Prediction of Airbnb Prices

Joseph Son (joeson@stanford.edu)

Introduction

What factors are most important in determining the price of an Airbnb listing? In order to answer this question, I developed a model to predict price based on the features of a listing.

This project includes a range of methods including linear regression, Lasso, ridge regression, random forest, support vector regression (SVR), and neural networks. Basic natural language processing (NLP) to analyze important words associated with different tiers of prices based on text from customer reviews and listing summaries are also presented.

Dataset

Dataset
The dataset used was obtained from InsideAirbnb and contains over 50k+ listings for NYC with information on property characteristics such as number of rooms and bathrooms as well as location.

Preprocessing
Features with frequent missing were removed and categorical features were used as factors. In addition, the features were normalized (by subtracting the mean and dividing by the standard deviation). Price was converted to log of the price to mitigate the impact of outliers in the dataset.

Models

Regression

- **Linear Regression** as a baseline model
- **Lasso** to perform feature selection
- **Ridge Regression** with cost function

$$J(\theta) = ||y - X\theta||^2 + \gamma||\theta||^2$$

- **Random Forest** with 50 trees and tuned to randomly select 5 features for each branching step
- **SVR** with a linear kernel was also trained
- **Neural Network** with a single hidden layer containing 10 neurons.

NLP
Supervised topic modeling was also performed by splitting the review and summary information into four pricing tiers based on the price of the associated listing.

Term frequency-inverse document frequency (TF-IDF) was used to find words occurring frequently in one price tier and not others. A term will receive a high weight if it's common in a specific document and also uncommon across all other documents.

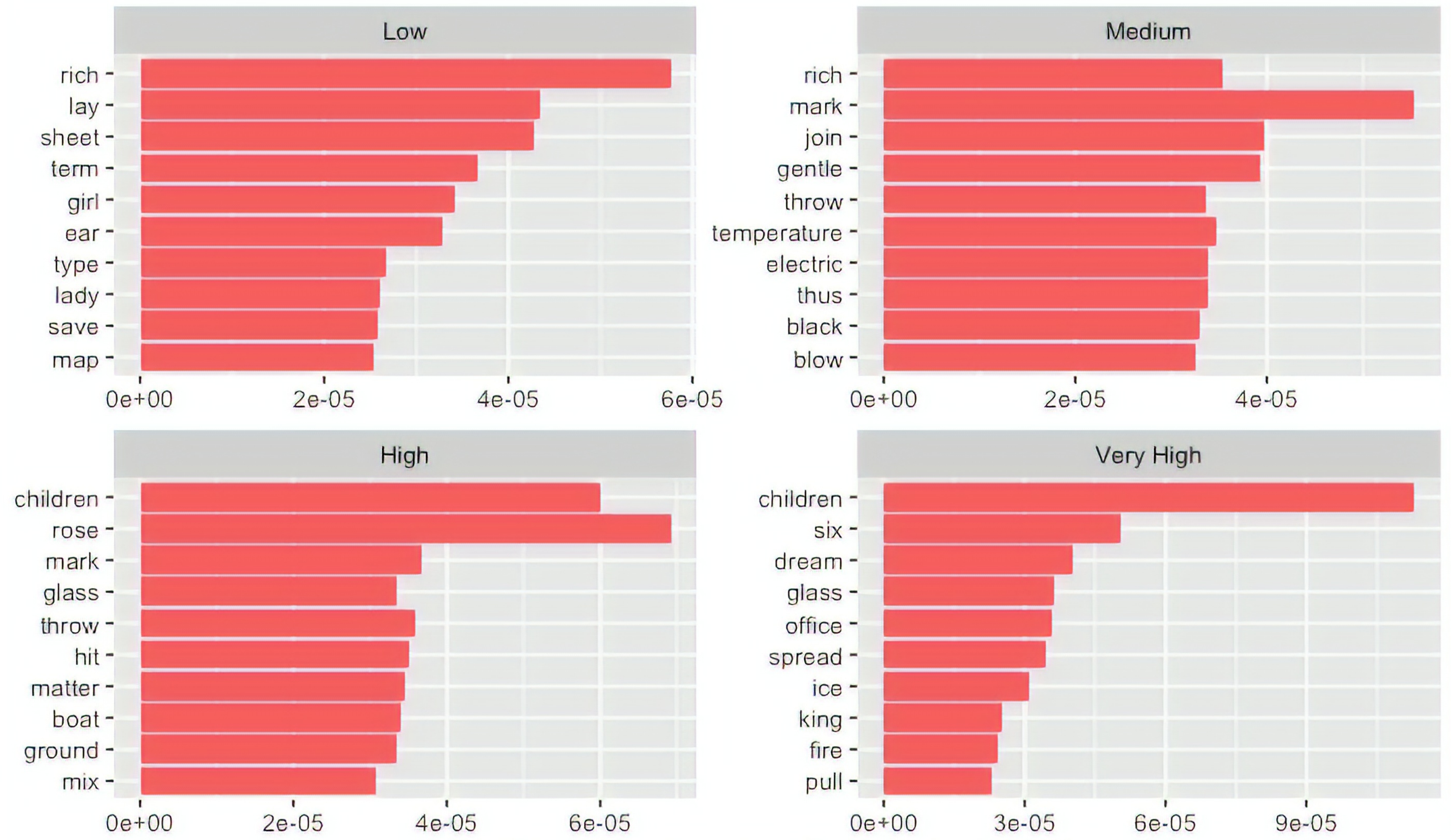
Regression Results

Model Name	Training		Test	
	MSE*	R^2	MSE	R^2
Linear Regression	1.18	0.6536	1.20	0.6293
Lasso	1.19	0.6408	1.21	0.6180
Ridge Regression	1.27	0.6065	1.28	0.5810
Random Forest	1.16	0.6987	1.17	0.6829
SVR	1.18	0.6457	1.20	0.6259
Neural Network	1.15	0.7068	1.18	0.6672

*MSE results were exponentiated to represent dollar amounts of squared error for each observation

NLP Results

Prior to using TF-IDF to perform supervised topic modeling, the dataset was split into four different price tiers roughly based on the quartiles of the price data.



TF-IDF on San Francisco Reviews by Price Tier

Conclusion

Based on the results, random forest was shown to be the most effective in predicting price. Although Lasso performed worse than linear regression, it was important in determining the variables most import to predicting price. Many of the coefficients remaining were related to the neighborhoods the listings were in. This makes sense given that the desirability and demand for listings in popular areas.

In addition to the results of the regression models, the supervised topic modeling on the review information provided for listings in SF were also quite informative. Pricier listings tend to focus on upscale aspects of the property and catering to larger groups like families.

Further work could be performed to improve pricing prediction using different models, tuning hyperparameters, and feature engineering. In particular, there is a large opportunity for further review analysis beyond the work performed for this project.