Seattle AirBnB Prices

Rob Palinic

Project Goal

What exactly is the right price for an AirBnB home? Each city and neighbourhood is a unique market, with hosts needing to decide their own pricing based on their own market research and gut instinct. Some hosts may be doing very well, with 100% booking rates, driven by below market pricing. Other hosts may be over priced, or taking actions that are not helping them charge higher prices. Analysis around this subject may provide some insight that would be of use to both groups of hosts.

Goals:

1. To create a pricing model which can be used to identify over and underpriced homes based on their characteristics.
2. Identify the importance of home size, location and host actions (as determined by reviews) on home prices

Data Sources and Processing

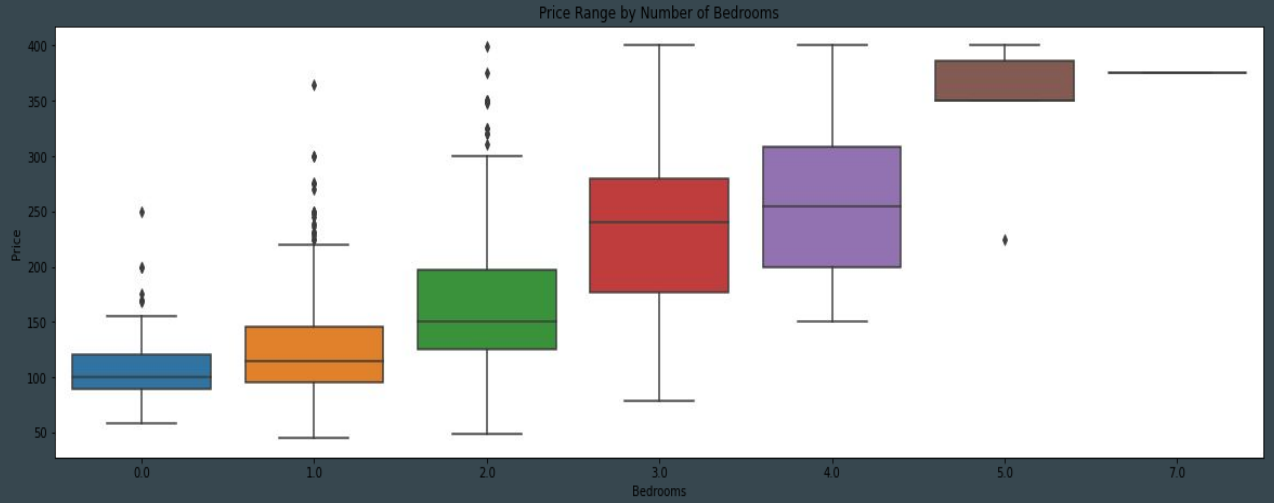
Data about listings have been acquired from the Kaggle website for Seattle, with a snapshot of listings from early 2016. There are two datasets: one for reviews, and one for listings. The listing dataset has around 2700 records and 92 fields.

These fields can broadly be broken down into seven groups:

1. Neighbourhood
2. Listing Description
3. Photograph URLs
4. Data Scraping Info
5. Host Info
6. Home Characteristics
7. Review Scores

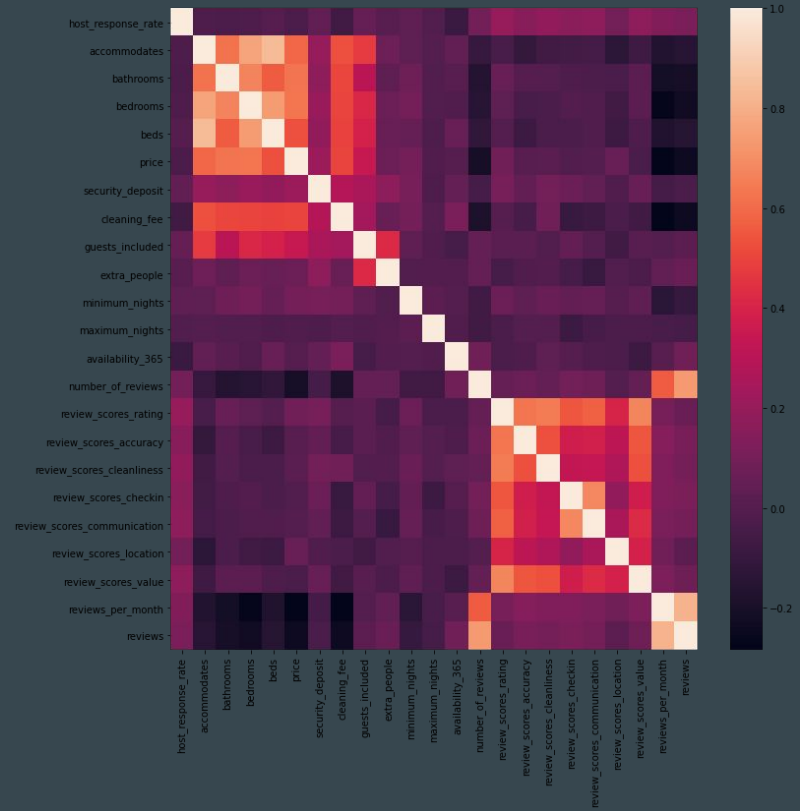
Many of the fields overlap or can immediately be dismissed as being irrelevant, such as info on when the data was scraped.

Exploratory Data Analysis - Home Size



There is a clear correlation between the number of bedrooms and the price range of the home, as is to be expected.

Exploratory Data Analysis - Numeric Field Correlations



Correlations between numeric fields show high correlations between size parameters (top left of the grid) and a high correlation between review scores (bottom right of the grid)

Price is also highly correlated with size.

Models Used

The purpose of this work is to predict prices, therefore regression models have been used. These include:

1. A dummy model to show the predictive of a simple guess
2. Normal Regression
3. Lasso Regression
4. Ridge Regression
5. Random Forest
6. XGBoost

Model Performance - Scoring

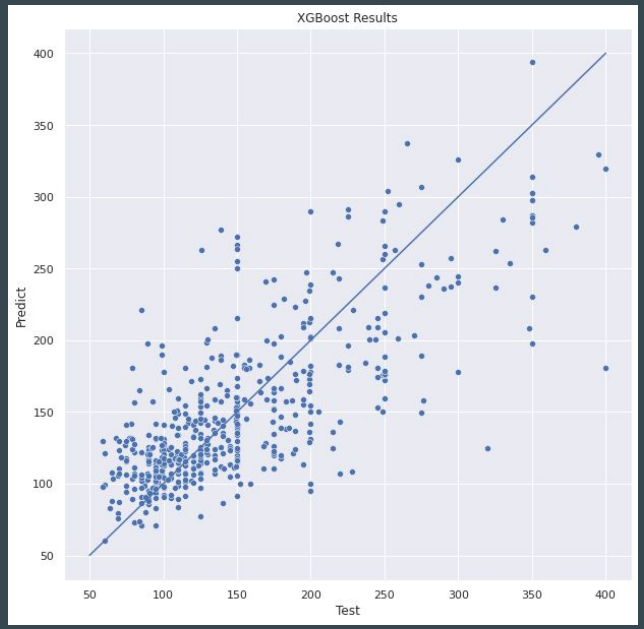
The Dummy Regressor performed poorly and did not explain the variability.

Every other model showed at least a 50% improvement in mean average error.

XGBoost had the highest feature explainability, and lowest MAE.

	Dummy Reg.	Linear Reg.	Lasso Reg.	Ridge Reg.	Random Forest	XGBoost
R^2	-0.01	0.53	0.54	0.53	0.52	0.57
MAE	51.12	33.96	33.89	33.95	34.79	32.86

Model Performance - Actual vs Predicted



This scatter plot of the predicted vs. actual values indicates a wide range, and is based on data from the XGBoost model (the best performing model).

At the low price end, the model over predicts, while at the high end, the model underpredicts.

Feature Importance

Linear Regression	Lasso Regression	Ridge Regression	Random Forest	XGBoost
bedrooms	bedrooms	bedrooms	bedrooms	bathrooms
bathrooms	bathrooms	host_response_time_within an hour	bathrooms	bedrooms
neighbourhood_group_cleansed_Downtown	cleaning_fee	bathrooms	cleaning_fee	cleaning_fee
cleaning_fee	accommodates	host_response_time_within a few hours	availability_365	neighbourhood_group_cleansed_Downtown
accommodates	neighbourhood_group_cleansed_Downtown	host_response_time_within a day	number_of_reviews	guests_included

The number of bedrooms and bathrooms is common across all models, while accommodates (the number of guests) is also important, as is cleaning_fee.

Several models also identify a downtown location as important.

Review scores are not important for any of the listed models.

Conclusions

All the models performed better than the dummy model, with at least a 50% improvement in scores.

The XGBoost model performed the best, but still had similar performance to other models. A regression model would be easier to explain and much quicker to run.

The model only explains a little more than half of the price variability, and can only be used for directional understanding and not for price recommendations.

Recommendations

1. Create a quality score, akin to a hotel's star rating. There are clear differences between listings based on uncaptured quality metrics. The model showed bias towards the topic and bottom price points, most likely driven by perceived quality.
2. Investigate supply and demand by size segment and neighborhood, and focus efforts on increased supply of tight demand unit types.
3. Focus effort on a greater understanding of price drivers through photographs and descriptions. Features may be uncovered which are important to renters that can actually be actioned by hosts.

Future Steps

a. Analyze Market Supply and Demand

b. Determine the effect of photographs on price

c. Analyze the effect of descriptions

d. Identify host pricing biases

e. Better feature selection