Project Luther - Craigslist Housing Rentals

Leon Zhou

Project Scope and Description

Using web-scraping tools, extract information from listings posted on the San Francisco Bay Area Craigslist housing forum. Scraping will be done primarily on the North Bay and South Bay sub-forums, and yield information such as numbers of bedrooms and bathrooms.

A linear regression will be performed upon the scraped data to generate a model aiming to predict the rental price of the property. This model will be valuable to those seeking to understand current rental market conditions in order to optimally price their own properties.

Methods and Analysis

The majority of data scraped was numeric, or were true/false values that could be encoded as such. The exceptions to this were the fields representing parking and laundry availability, as well as the type of property to be rented. The property type contained 9 unique values, and was converted into a one-hot encoding.

In order to limit the dimensionality and number of features to regress upon, laundry and parking were encoded as numeric values, with higher numbers representing a more desirable or comprehensive amenity; for example, street-parking only was encoded as a 2, while an integrated or attached garage was encoded as a 6.

While this serve its purpose in simplifying the model, in hindsight, it raised a different problem. The differences between each level were spaced in increments of one, but the difference in added value to the rental may not have been equally spaced - a difference between “no parking” and “street parking” may have been more valuable than a difference between “detached garage” and “attached garage,” potentially requiring additional terms to account for.

The combination of features and encoded property type yielded a total of 22 columns. A train test split of 70%/30% was used. To simplify the model and avoid overfit, a Lasso variation of the ordinary least squares regression was performed. This eliminated six features, as well as some of the sparsely-represented property columns.

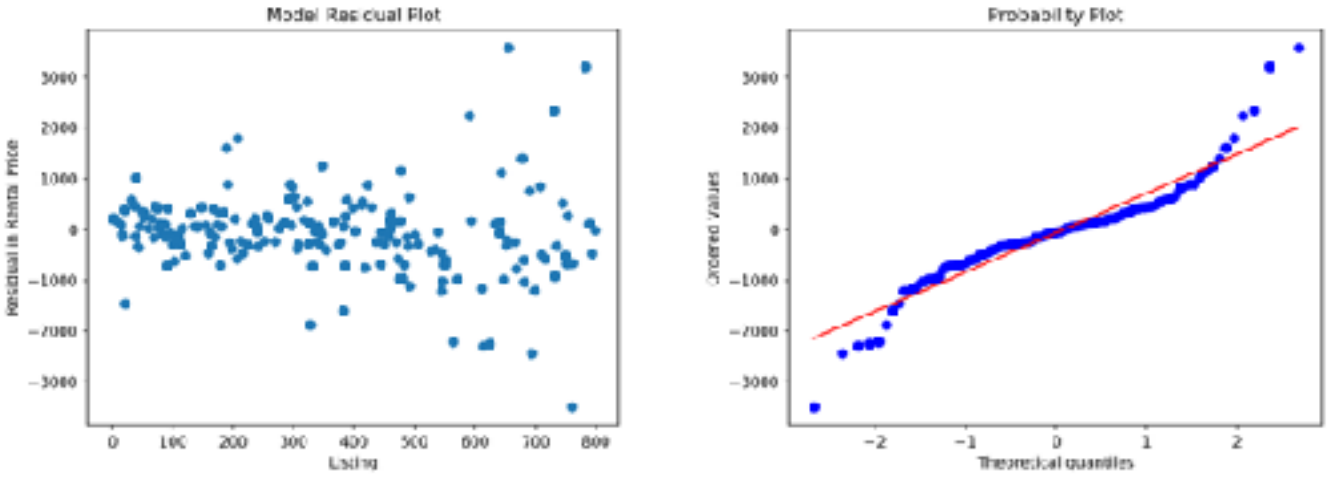

Figure 1. (left) Residual of predicted rental price. (right) q-q plot of residuals.

This trained model was validated on the unseen test set, and achieved an R^2 value of 0.658. A q-q plot of residuals showed a mostly normal distribution, though some waviness was observed. A very small amount of heteroskedacity seems present, suggesting space for improvement in the model.

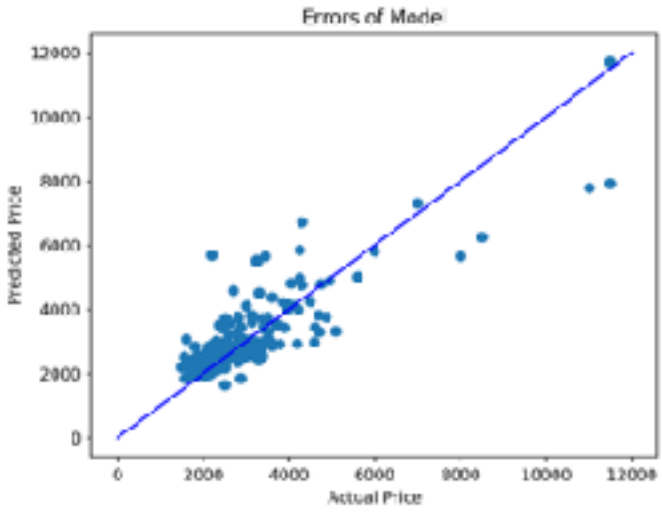

Figure 2. Actual price plotted against predicted price, with a 45 degree guide line.

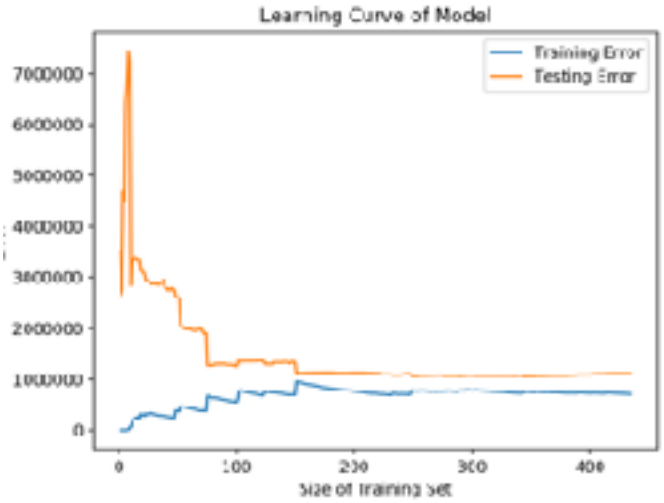

Figure 3. Learning curve of model, taken in increments of one.

Future Exploration

The model, in its current state only distinguishes geographically between North Bay properties and South Bay properties. Some of the unexplained variation could be due to treating each region as a monolithic entity, when socioeconomic and demographic properties could vary greatly neighborhood to neighborhood. The data for doing so is already present; a general address and neighborhood is listed on the majority of housing posts.

A second useful addition, should the data be available, would be examining if the listed properties were successfully rented, and the price they were rented at. It could be that many of the most expensive properties had no takers, and the actual demand curve is lower than one might be led to expect merely looking at the supply-side of things. Incorporating this data may also make features such as images or words in the post more significant, as the presentation of the listing could conceivably have an effect on renters.