Beating the Zestimate: Predicting Home Price Values in Seattle

I recently set out to build a model to predict home prices in Seattle, WA. Granted, this was not an easy task. With all of the different features any particular home may have compared to another, real estate can seem tough to value. Still, any home has features that undeniably make it more, or less, valuable - e.g. square footage.

For most homeowners, the Zillow "Zestimate" is the closest thing they have to an objective valutation of their home. Many people use the Zestimate as a measure of their home's equity, and by proxy their net worth. Given the data set we had availble, my goal was to see if it was possible to meet, or even beat, the vaunted "Zestimate" in the city of Seattle.

Data

We started with a dataset of homes that were sold in 2015 from Kaggle.com. Features included:

*Price the home sold for,

*Square footage,

*Lot size,

*Number of bedrooms & bathrooms,

and over a dozen over variables. This data set gave us a good head start, but the objective features of a home don't tell the full story. Our next step was to do some additional feature engineering that would help us better predict what price each home was sold for.

Feature Engineering & Web Scraping

Our first step was to query the Google Maps to get the closest possible addresses from latitude/longitude info provided in the original dataset. Then, we used these addresses to get a list of the exact census tract where each of the 21,000 houses was located. We then took it one step further, using Census Data from the 2015 American Community Survey (ACS) to get Median Household Income for each row in the data set.

We also used web scraping to further the accuracy of our data set. We chose to scrape data from Walk Score (www.walkscore.com), since it is a well-known and respected metric to describe the walkability of a given home's neighborhood. We hypothesized that walkability would be a good predictor of a home's value - in short, people are willing to pay more for homes that are near the places where they work and play.

In addition to the prototypical "Walk Score," the site has other scores we thought would also be useful. Along with a "Bike Score" and "Transit Score", we thought that a metric of particular importance would be the Crime Grades that the site offers, which are based upon their study of personal and property crime histories for each area. After adding these to our data set, we felt we had enough information to begin our analysis.

Methodology & Results

As noted previously, we hypothesized that Median Household Income, Crime Data, and Walkability would be important independent variables that would make our model more accurate. Since this data was only available for the city of Seattle (rather than all of King County), we chose to focus on homes in Seattle that had Walk Scores and Crime Scores.

We quickly determined that the best dependent variable for our analysis was the log of home price, since home price itself did not follow a normal distribution. We then modeled a basic, ordinary least squares regression (which was cross-validated using train/test split) to get a baseline. This model produced an R-squared value of 0.742 - a good starting point..

Building upon our initial results, we performed feature scaling on our data, used scikit learn's standard scaler in order to make our data more uniform. We did a train/test split on our data in order to cross-validate our results, then fit the data to a lasso model with built-in cross-validation in order to eliminate redundant features and reduce complexity. This analysis gave us a higher R-squared value, and identified some features that were redundant and we were able to drop. Those features were rather surprising - it turns out that the size of a home's lot does not seem to matter a whole lot in determining its price, at least in Seattle! The more important variables were things like the condition and grade of the home (as measured by the King County Assessor's Office), and good old square footage. It turned out that Walk Score was rather important, too - it rated as one of our top variables.

After removing the variables which did not add to our analysis, we fit the newly cleaned data to a Ridge regression model (also with feature scaling), and added second-degree Polynomial features to tease out additional interactions within the data. Polynomial features make sense in this case since housing prices can often exhibit non-linear characteristics, especially on the high end - in fact, they can go up exponentially!

Final Results

In the end, after testing multiple different parameters for our linear regression model, we ended up with a Lasso model which produced an R-squared of 0.781 (train) and 0.742 (test), with a mean squared error (MSE) of .046. In layman's terms, our model was able to explain nearly 80% of the variability in our data - a surprisingly good result for an entire region's housing prices!

Next Steps/Future Work

Given more time and data, we'd love to do a few additional things. For one, we'd like to be able to add data about the quality of schools in each neighborhood - as anyone who's been in the market for a home knows, the quality of schools in the area matter a lot! We'd also like to create separate models for homes that are truly "high end" of the market - those homes seem to be less tied to the traditional metrics of valuation. Finally, we'd love to create a web app to show off the results of our model, and allow anyone who lives in Seattle to enter the characteristics of their home, and get our model's estimate of how much their home is worth.

Beating the Zestimate...?

Our analysis of the cost of homes in Seattle was an excellent start. Zillow admits that their "Zestimate" is not terribly accurate in Seattle. In fact, they give themselves 2 out of 5 stars for accuracy - not exactly a vote of confidence. They also admit that their estimates have a median error rate of about 5%. Our model came surprisingly close to beating this metric - but we're not ready to declare victory just yet. Zillow has heaps and heaps of proprietary data about the housing market about both sides of the market - buyers and sellers - that we simply aren't privy to.

Still, with more time, I think we could improve the Zestimate's accuracy. Our model was surpisingly accurate, and beating Zillow is only a matter of more time, and more data. The gauntlet has been thrown down!

Thanks for reading. Until next time!