Predicting Housing Prices in Ames, Iowa, using Multiple Linear Regression

Presented to the Chief Information Officer at MyHomeSearch

Problem Statement

MyHomeSearch is an emergent competitor in the online real estate data industry and is interested in increasing its web traffic in relation to competitors like Zillow and RedFin

Research questions:

- What house features could be relevant to a prospective home buyer not yet listed in major websites like Zillow?
- How reliable are those features in predicting the prices of future unknown homes?

Data Sources

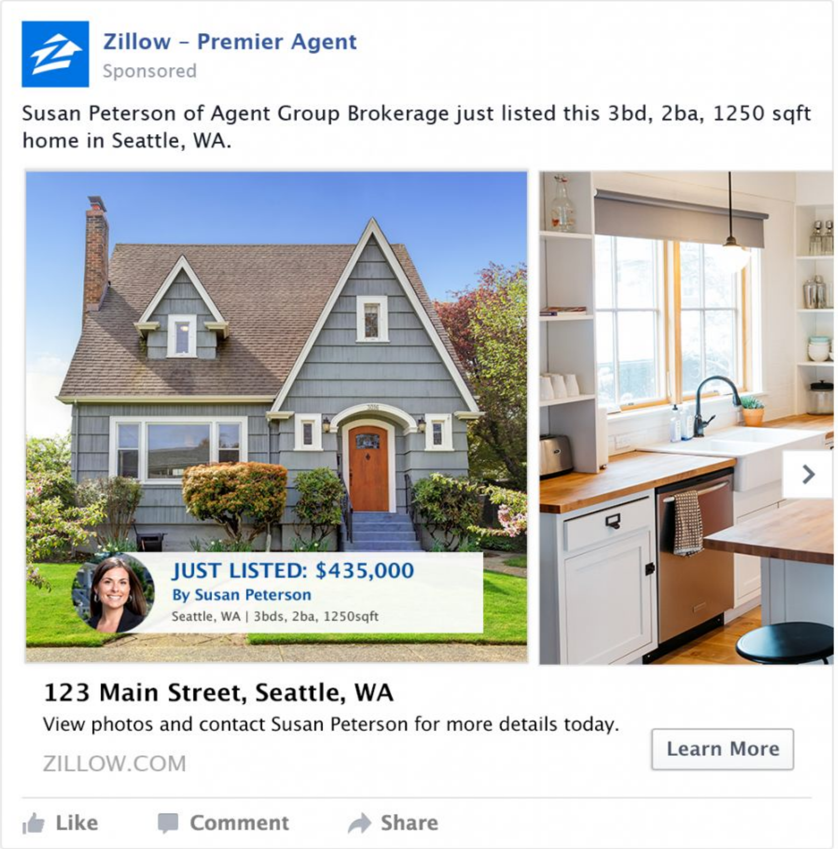
- Housing characteristics data from the city of Ames, Iowa's Assessor's Office for residential properties sold between 2006 and 2010
- The data set included 80 features related including information on lot area, number of bathrooms, sale price, and more
- Variables were of different types including continuous, discrete, nominal, and ordinal variables
- The data set represented a total of 2,929 houses

Methodology

- Identify what features the data set included that Zillow does not list online

- Describe and visualize the relationship between each feature in the dataset to the sales price target variable and select a subset of variables to estimate the sales price of a house using a correlation matrix heatmap and multiple linear regression

- Predict the sales price based on selected features and validate the predictions against a test data set of unknown and unseen data to a machine learning model



Zillow – Premier Agent

Sponsored

Susan Peterson of Agent Group Brokerage just listed this 3bd, 2ba, 1250 sqft home in Seattle, WA.

JUST LISTED: $435,000

By Susan Peterson

Seattle, WA | 3bds, 2ba, 1250sqft

123 Main Street, Seattle, WA

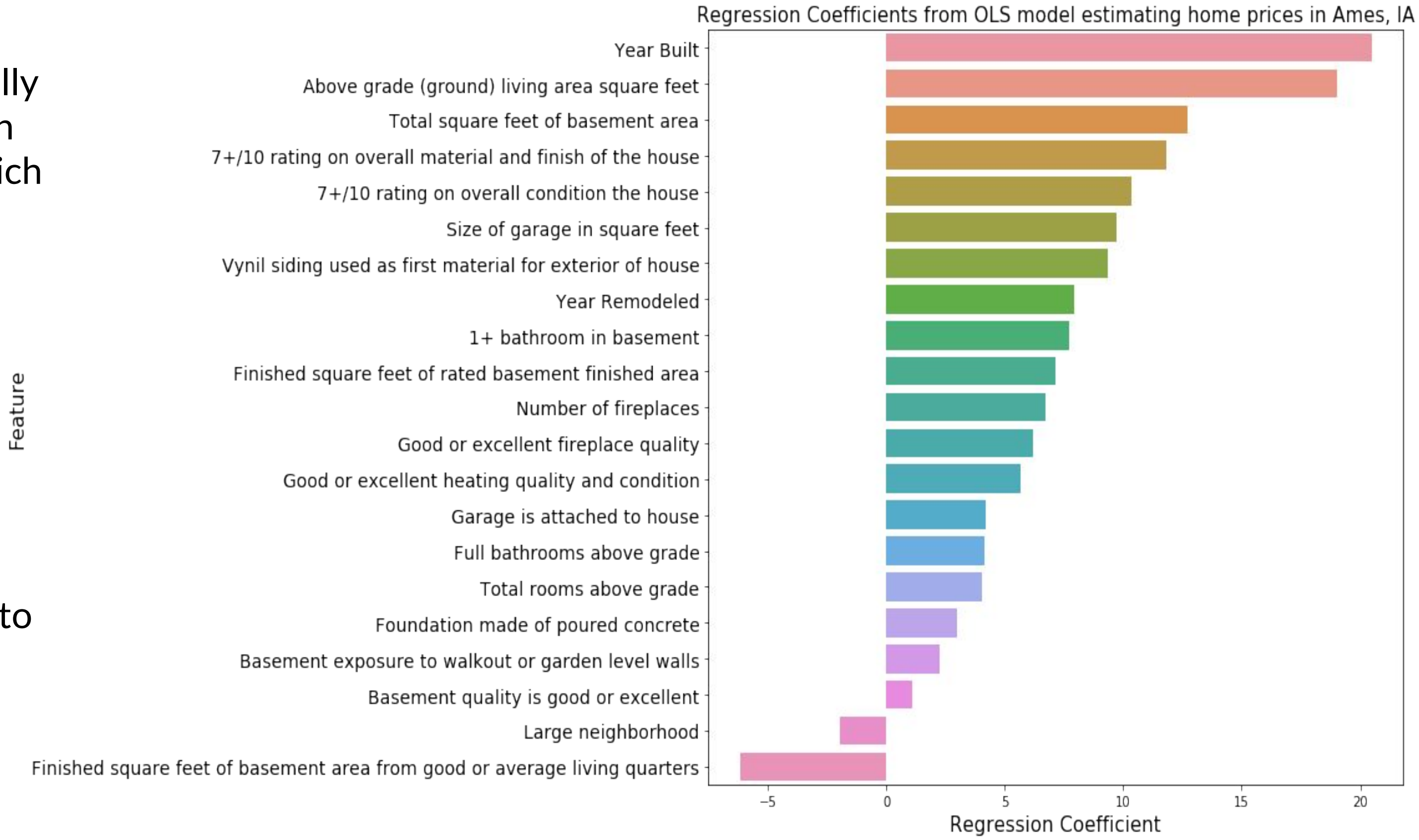
View photos and contact Susan Peterson for more details today.

ZILLOW.COM

Primary Findings

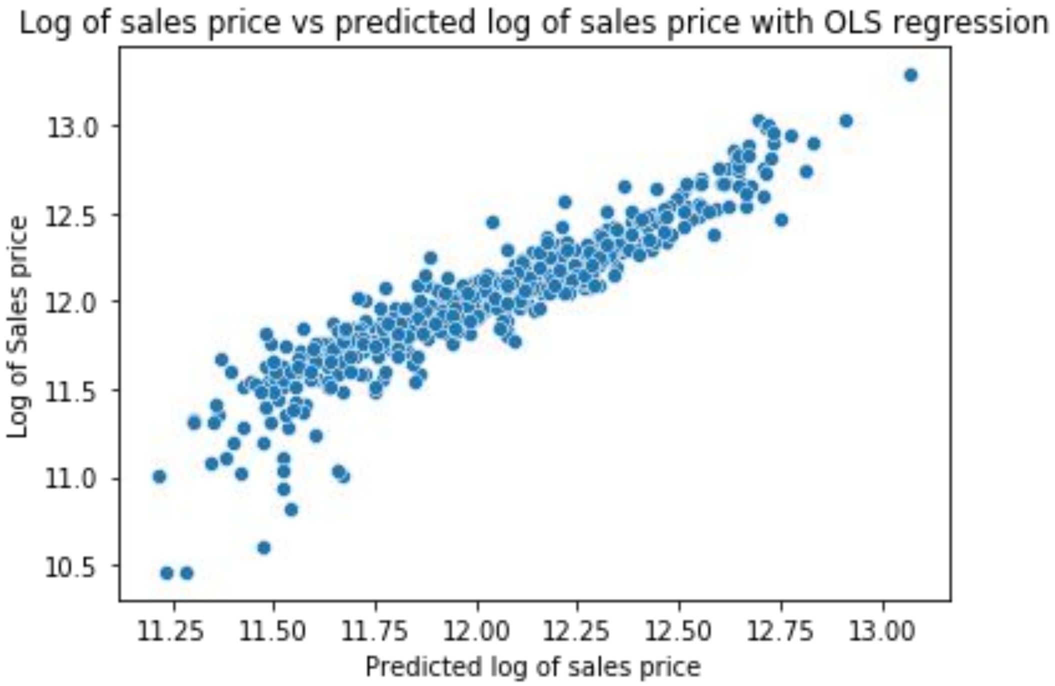
21 features had a statistically significant relationship with the sales price, some of which are not listed by Zillow

Houses who were rated as good, very good, excellent, and very excellent in the overall condition and **materials** of the house, on average, were more 12% more expensive compared to houses with lower ratings



Primary Findings - Cont'd

- Multiple linear regression was used to predict the sale price of a house were on 21 features
- OLS, Ridge, and Lasso regression models were fit to the data to estimate the regression model coefficients
- Overall, the model explained around 83% of the variation in the sales price in the training data set and 87% in the testing data set using the OLS model
- The model did not perform as well toward the extremes of the distribution of the log of sales price



Conclusions & Recommendations

- Consider publishing house facts that rely on subject matter assessments of specific areas of strength or need for a house
- This information may be particularly important for individuals who want to buy houses in poor condition to improve them at a profit
- The prediction of sale price should be improved to reach an R-squared value of around 90-95% to ensure as much accuracy as possible
- An analysis should be done to evaluate and understand the accuracy of Zillow’s Zestimates and use this information as a baseline to measure model’s reliability