Anthony Maimone
King County Housing Project
19 November 2020

Objective

The main purpose of this project is to predict the price of houses, as well as identify the factors that are most impactful in determining a house price. Additionally, I want to answer the specific question below.

I am looking to buy a 4-bedroom 3 bath house in the Eastside area, specifically Kirkland. If I want the house to be good grade and condition (8 for both), about what price can I expect?

The Data

The data I used was the provided King County housing data. Additionally, I pulled data from the web that listed every zip code in King County with the city name that each zip code is within. The reason behind this was so that the data could be aggregated by location more easily by having a smaller number of groups, rather than the large amount of zip codes. To add in the city column, I used Power Query within Excel to join the two query tables on the zip code column. This appended another column to the data set with the name of each city that each house is within. I choose to do this in this way so that I would not have to manually assign a city to each zipcode in the dataset.

I started with some initial exploratory data analysis, to gain initial insights on the data. I grouped by city and then calculated the average house price for each city. Medina, Mercer Island, and Bellevue appear to be the most expensive cities in King County. I then created 3 visualizations to confirm what key factors affect the price of a house. Two charts display positive correlations between house square footage and price, as well house grade and price, indicating that an increase in either square feet or grade will also increase price. The last chart shows house conditions and price. Though the house condition does clearly affect price, the correlation is not as strong as the other two variables, as evident by the general increase in price as condition grade level rises shown in the third chart. This condition vs price visualization does show a few outlying data points within conditions 3 and 4, however overall there doesn't seem to be many outlying points so they should not contribute a large amount to the error.

To simplify the models, I chose to focus the data on cities that are typically considered to be on the Eastside of Seattle. This was done by filtering to a specific list of eight cities. I then needed to created binary columns to represent the cities so that the cities could be used within the models I would make. After creating binary columns, I removed columns that were not likely to affect the house price such as date and id. Finally, the data was split 80-20 into train and test sets, and preprocessed to remove any near zero variance columns for better model performance.

Method

I created two models to predict home prices. Based on the initial findings about how square feet and condition and grade affect price, I decided to make linear models. The first model was a simple linear model with all features used. I opted to use mean absolute error for the accuracy metric, since it would be more useful to measure the actual cost deviation in predicted value of a house. This first model resulted in a mean absolute error of $136,862.7, which is not a small number to be off on when considering buying a house since that can easily push someone outside of their maximum budget. I did create a variable importance plot, so that I can do a high level look at which variable are more integral to the model's performance.

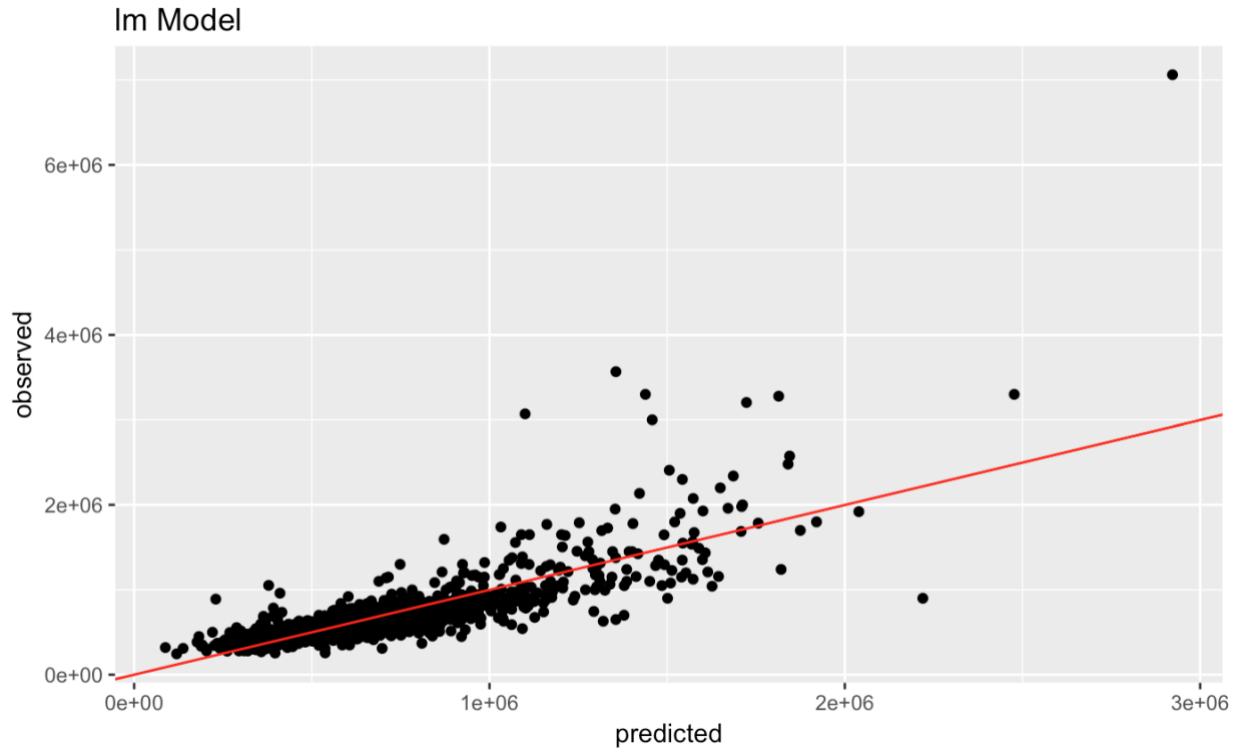
Unsurprisingly, the cities, zipcodes, and square feet of each house proved to be the most significant variables.

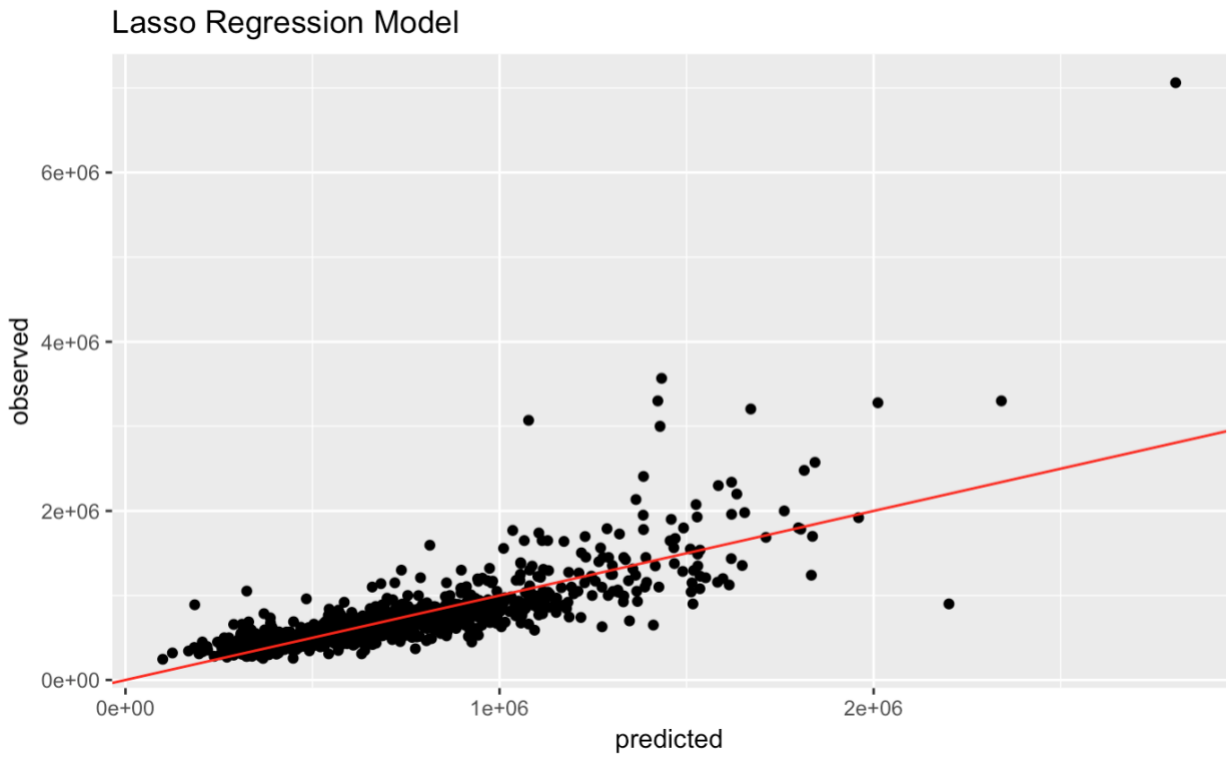
The second linear model I made used lasso regression. This was beneficial because I was able to test different tune length and cross validation parameters to help reduce overfitting the model. With this model, I only included the variables that had a high variable importance from the variable importance plot in the first linear model. Again, I chose mean absolute error for this model for the same reason as the previous one. The lasso model had a slightly improved error of $136,395.6, which is a difference of $461.7. While this reduction in error is a substantial improvement, relative to the total error is it not as major of a difference in terms of the overall predictive accuracy of the models.

Visualizations

I did provide some visualizations in my exploratory data analysis which as mentioned earlier, showcased the correlations between square feet and price, as well as house condition and grade and price. These all confirm that each variable does in fact play a considerable role in determining the price of a house.

I created a visualization of the errors for each model to get a visual understanding of the predictive accuracy of them.





Both these charts show that each model performed similarly and that generally, the predicted values follow the model's approximation. However, there are clearly strong outliers and given the high mean absolute errors that occurred, these clearly do affect the models' accuracy.

Conclusions

Comparing the two models, the linear model with lasso regression had a slight advantage on the simple linear model since the lasso regression allows you to adjust more parameters such as cross validation. The lasso model also benefited from using just the most significant variables in determining a house price such as location (city, zipcode) and available square feet.

The major conclusion that can be drawn from this data is that the location of a house (as in what city it is in, but also even more specifically what zipcode) is typically the biggest factor to affect a house's price. Other variables such has square feet, number of bedrooms and bathrooms, as well as house condition and grade play a large role as well in determining a house's price, however location seems to be the biggest driver.

Lastly, I created my own data point to test the lasso model on. I chose the lasso model since it performed slightly better than the other. The purpose of making this single row data frame was generate the type of house I am looking for in my question in the objective statement. I input a 4 bedroom 3 bath house in Kirkland with 1800 square feet and condition and grade 8 into the lasso model, and it predicted a price of $566,071.7. This is likely an underestimate since the average house price in Kirkland was earlier identified as $646,374.2, and a 4-bedroom house is typically larger than most houses. This example shows how the error can skew the predicted values of price.

Overall, I would say that the models performed alright, but not good. The error was high and a difference in price of that much could easily affect someone's ability to purchase a house.