Reddic Housing LLC

Presented by
Jeff Mayah
Kavyn Abel
Bridger Hackworth
Jacob Allen
Paul Weinberg

Machine Learning Model Summary

To best predict the selling prices of housing in the Seattle area for Reddic Housing LLC we have decided to create a regressor model. We chose a supervised learning approach as we have historical data with the selling price of homes that have previously sold in this area.

I added an age column as well as a yard space column calculated from our existing data, as well as the proportion of the house in square feet compared to its nearest 15 neighbors. After normalizing the data we were able to split it into training, validation, and testing sets so we could train, tune, and test our model.

We were able to originally use some external data from King's County open data website that helped our model significantly. This data allowed us to create a crime rate by zip code feature. Disappointingly however, due to some Google Collab malfunctions, our highest performing model was deleted last minute. When attempting to reconstruct the model using the external data once more, we were unable to achieve the same results. In fact, our model that didn't use the external data was performing at a better rate than our attempt at the reconstruction. With more time we could likely improve our model to the same performance that it was at before using the crime rate data.

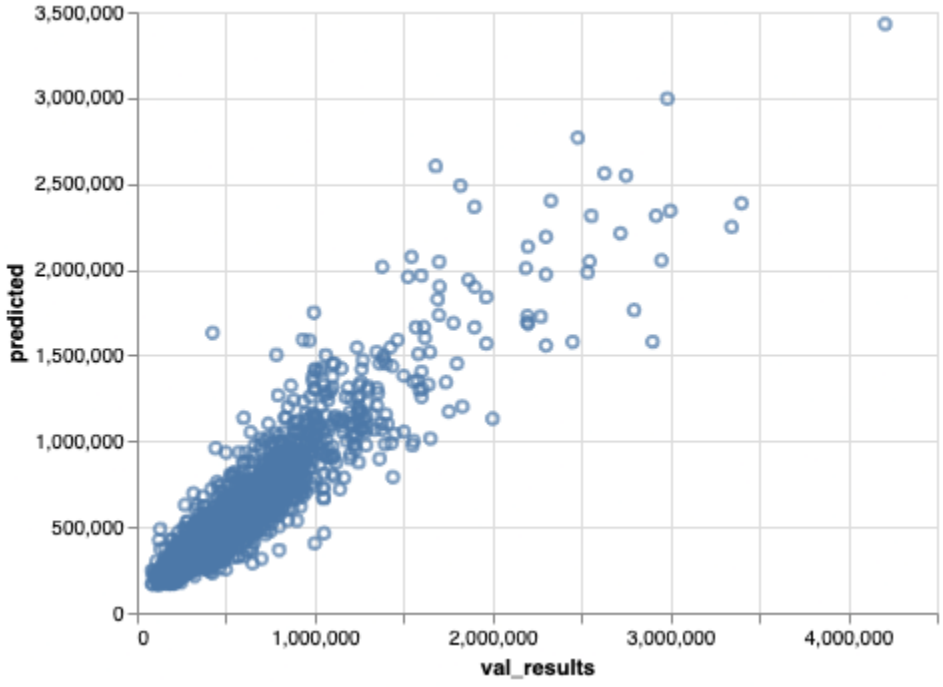
We chose to use the XGBRegressor model. This model creates an ensemble of boosted trees. This means that one tree is used, evaluated, and the errors are used to train the next tree to get us closer to the truth. When we put all the trees together, that is what makes it an ensemble. As we tried different features and parameters we were able to increase the

effectiveness of this model. We adjusted the max depth of each tree, that learning rate(being how much of a change the model will make after each tree is built, and the subsample to help prevent our model from overfitting the data. These parameters were tuned with our validation set, and then we assured that our model was still effective by testing the test set.

To best evaluate our model we went to the RMSE and R squared values. The RMSE helps us know on average how far our predictions are from the actual house prices. In using the testing set for our model our RMSE came out to be $130,473 showing that on average our model is $130,473 dollars away from the actual price of the home.

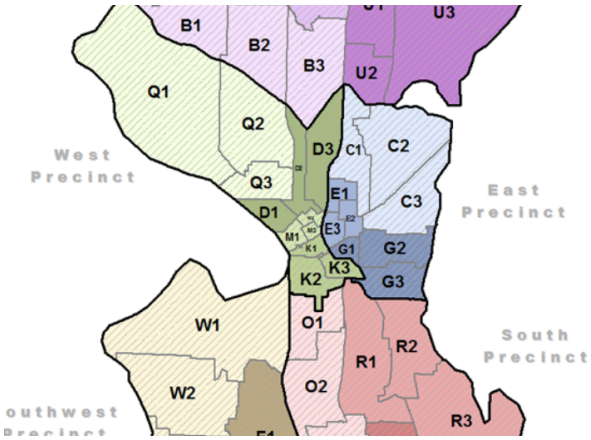
Additionally, we used the R squared metric to see how well our model fit the data. This metric is from 0 to 1, the closer we are to 1 the better our model fits. After running our testing set through our model we calculated an R squared value of 0.86. This is fantastic and means that our model fits the data well!

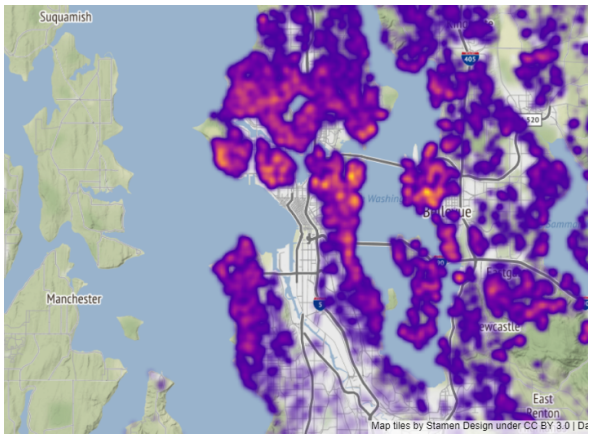
To help you see how well our model fits here is a graph of the actual prices from the tested set in relation to the amounts that our model predicted.



I. Adjusting Price of Low Income Areas

Assuming that the crime rate will be higher in areas with low income, we decided to make a column that correlates the crime rate and the price. We added this information to our model to adjust the prices. The graph on the left shows the crime rate based on each district in Seattle. The darker the section, the higher the crime rate. The map on the right is a heat map showing the concentration of home prices. The more yellow it is on the map, the higher the concentration of house prices. You can see there is an inverse relationship between crime and price concentration. This suggests that people are selling their home for a lower price when crime is high in an area.

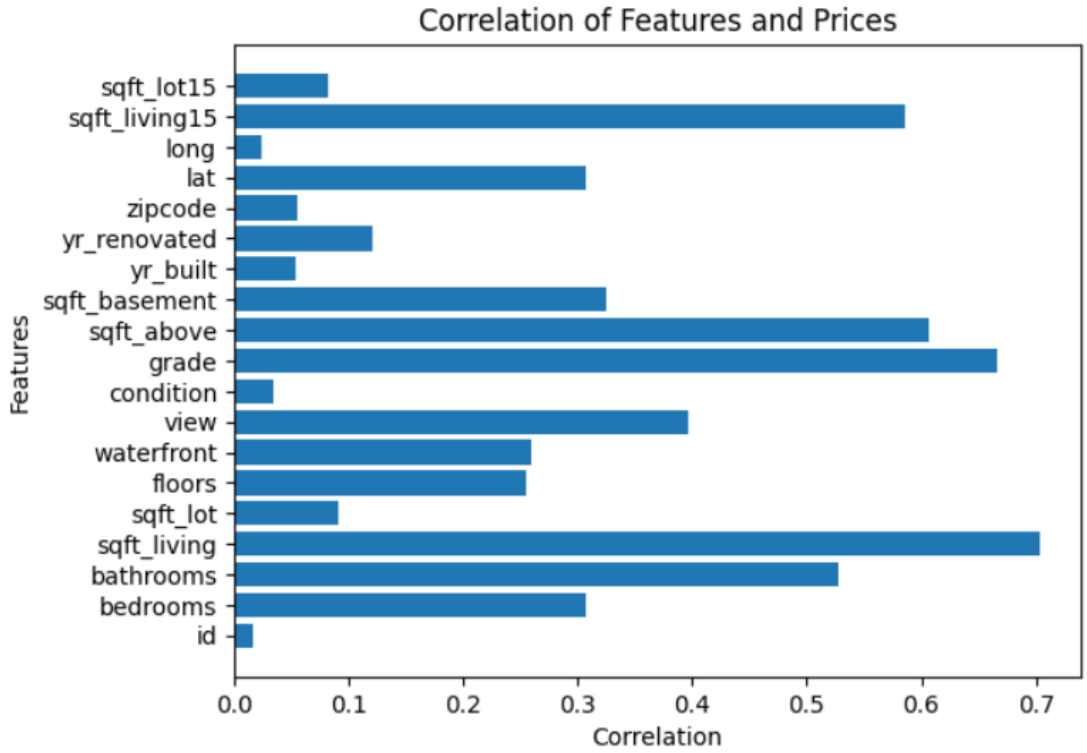




Lorem ipsum dolor sit amet, consectetuer adipiscing elit.

II. Features that affect the house prices the most

To determine the features that will influence the prices the most we decided to find the correlation between the prices and the different property types. The graph below shows the result of different correlations. A brief examination of the graph shows that space and quality exert a major influence in the prices than any other feature.



III. **Python Notebooks**

Below are Github Gist links to the notebooks we used during this case study:

Primary Notebook:

https://colab.research.google.com/drive/1P_Qoz66mmBY0Q-15VILzKrpw_357x43o?usp=sharing

Secondary Notebook for machine learning model:
https://colab.research.google.com/drive/1P_Qoz66mmBY0Q-15VILzKrpw_357x43o?usp=sharing

https://colab.research.google.com/drive/1X59VWyzxfj_DlEI2PwEyhM9sjkLeZmLU#scrollTo=2_qKA5QzrUt4

Correlation:
https://colab.research.google.com/drive/1bkAQN_plHJJJDCcvw9xuOqEVYgheuzpX#scrollTo=mA0HPVmIBT4C