Executive Summary

Module 3

presented by
Ben Jacobs

Brock Hoskins

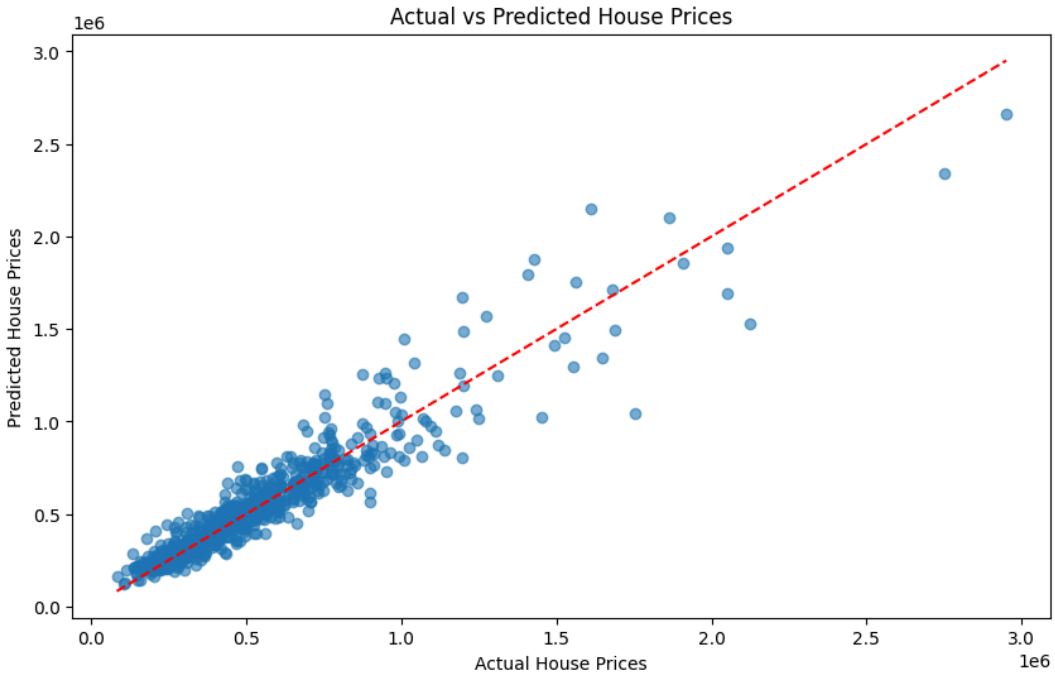
Conner Lacey

Eli

Joshua Ludwig

Vannah Pyatt

Model



Results

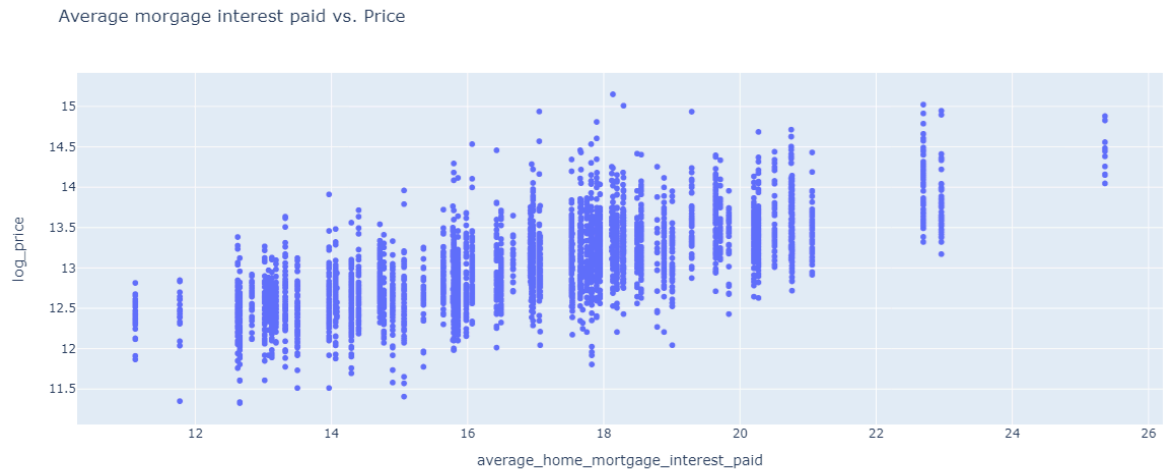
The model we decided on to predict the house prices was a gradient boosted tree. What this does is start the model out with a simple decision tree. This first tree in this model likely chooses the average values. After it runs through each feature, the model will then look at how far off the predicted values were from the actual values. It then makes a new tree after learning from these residuals from the columns we trained it to use. It does this until it gets as accurate as it can be.

After we had trained our model and seen the predictions of this model, we were able to get an idea of how closely we were predicting the actual value using the Root Mean Squared Error (RMSE). We chose this metric to evaluate our model for a couple of different reasons. The main one being is that it is an easy way to interpret the results because it keeps the predictions in the same units as the actual values. For housing prices in particular, there can be big differences in house prices from one house to another. In other words, the error can be quite large. The RMSE helps because it penalizes the large errors more than the small ones to better help us understand what we need to do better. The RMSE that our model was able to produce is $98,514.23. This tells us that when we predict a house price with our model, we can say with certainty that the actual house will sell within that error of $98,514. As you can see in the graph above, the error is on average smaller for homes that cost less and larger for homes that cost more. Below we discuss some ways we would combat this issue in the future.

The Data

We discussed the need to make sure we are **correctly identifying prices of homes that are in lower income areas**. To address that concern, we accessed tax information for each zip code from the IRS website. That data includes the adjusted gross income (AGI) in each zip code, the amount of salaries and wages in AGI, and even the home mortgage interest paid in that area. We used this information to find averages for each zip code to use in our model. Below is an example of the correlation between the price

(In the graph it is the log of the price to be able to see the correlation more clearly) and the average home mortgage interest paid by zip code. We can see a strong correlation, which helps us identify the areas with lower income housing.



Another thing we did to address this concern is that we found the area that has the highest prices of homes and treated that as an epicenter of sorts. We calculate the distance of homes from this epicenter to use as another indicator of the income level of the area. In the graphics below we can see why this metric is useful. In fig 1, the average price of homes (measured by circle size) decreases the further their distance from the center of Seattle, generally.

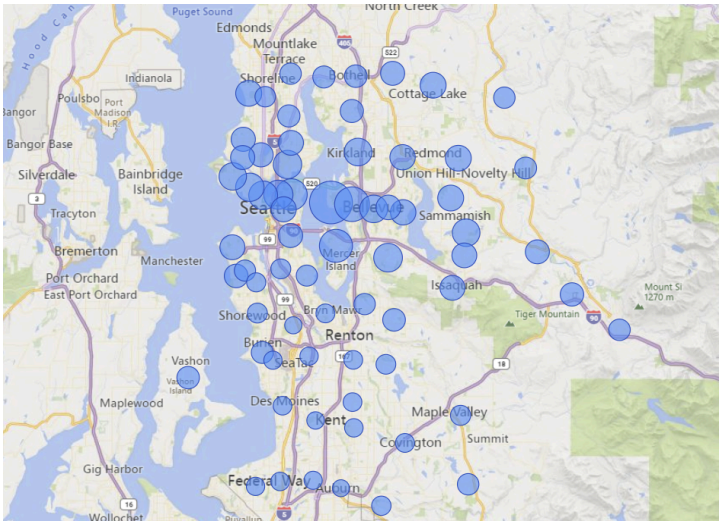
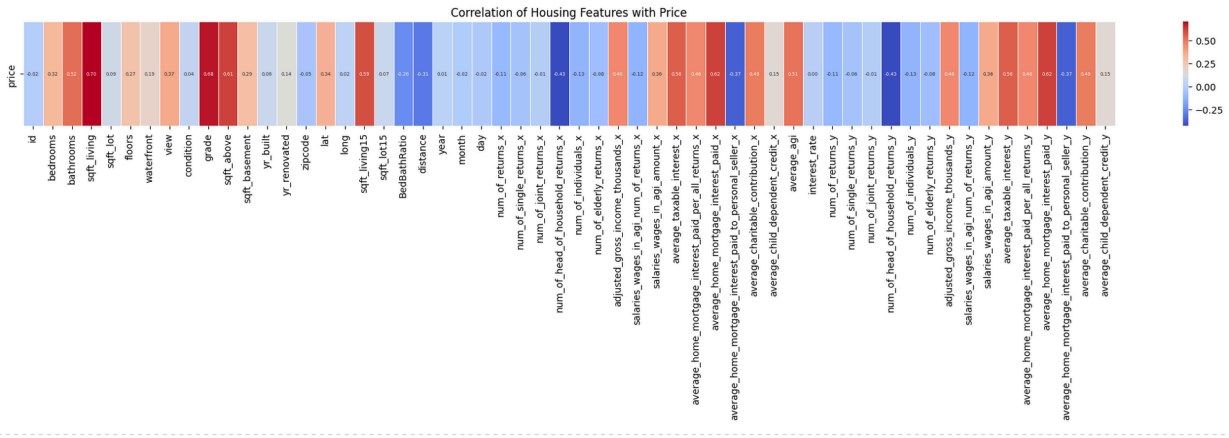


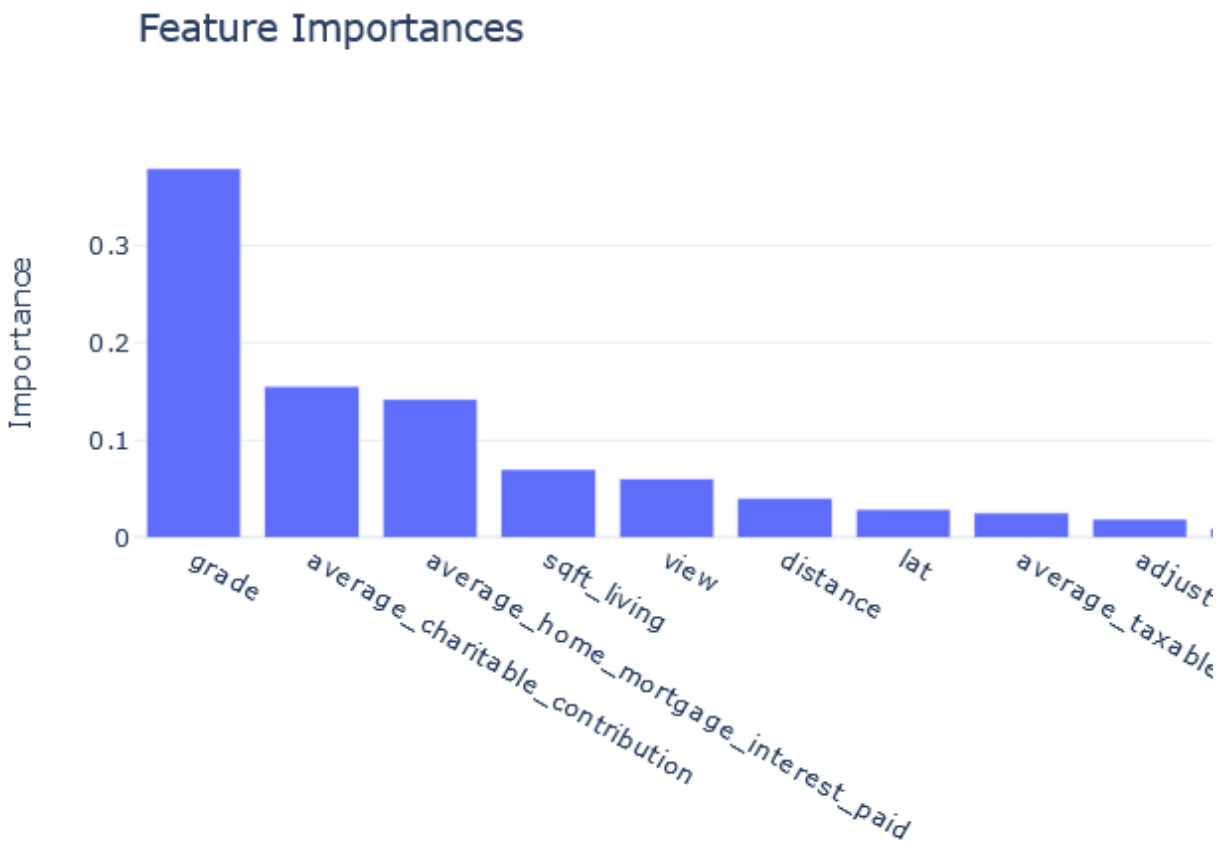
Fig 1 (Average Price of homes by zip code)

In addition to the tax information data, we also added data about mortgage interest rates each day. Since prices of homes fluctuate based on interest rates, we can use this additional data to make sure over time, our estimations change along with the interest rates. There is more information about that below.

This heatmap shows the correlation of features with price.



Notice how many of the highest correlating features also end up being of high importance in the model, such as grade and living area (sqft_living):



These insights helped us better tune the model.

Model Evaluation

We chose to use RMSE as the unit to determine the effectiveness of the model. This is because RMSE as a unit equates fairly well to price, and price ranges are how we are trying to evaluate the model.

If we did not add the additional data and metrics to our model that is called a vanilla model. The vanilla gradient boosted model scores an RMSE of around 133,000, meaning each guess is on average about $133k off from the target price. The final model had an RMSE of about 98,500, which is about an improvement of 34,500 or about 25.5%.

Note that this is the model on general housing. See the action items for ways that this score could be further improved.

Methods

To find the best possible combination of elements to include in the model, we created a loop that iterates through every feature in the data frame. For each iteration, a copy of the data frame is made without the iterated feature. That data frame is then run through a performance test. Once the loop is completed, the results of every iteration are compared. The removed feature from whichever iteration performed the best is permanently removed from the dataframe and the process is repeated until only one feature is left. We then single out the combination that performed the best.

The features used in the model were the following:

```
['bedrooms', 'sqft_living', 'sqft_lot', 'floors', 'view', 'condition',
'grade', 'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated',
'zipcode', 'lat', 'long', 'sqft_living15', 'sqft_lot15', 'distance',
'num_of_head_of_household_returns', 'num_of_elderly_returns',
'adjusted_gross_income_thousands', 'salaries_wages_in_agi_num_of_returns',
'salaries_wages_in_agi_amount', 'average_taxable_interest',
'average_home_mortgage_interest_paid',
'average_home_mortgage_interest_paid_to_personal_seller',
'average_charitable_contribution', 'interest_rate']
```

Action Items

- One thing we want to do to further address the issue of correctly identifying prices of homes in low income areas is to use the location of the home (longitude and latitude) to find the average price of the nearest homes we know. This would also allow us to better identify the lower income areas and ensure that our predictions are more accurate using the data that we currently have on homes in the Seattle area.
- The model can be modified to better fit higher cost housing, or lower cost housing. Here is a graph showing the index of the cost, and the price. We see a sharp spike around $1.8 million.

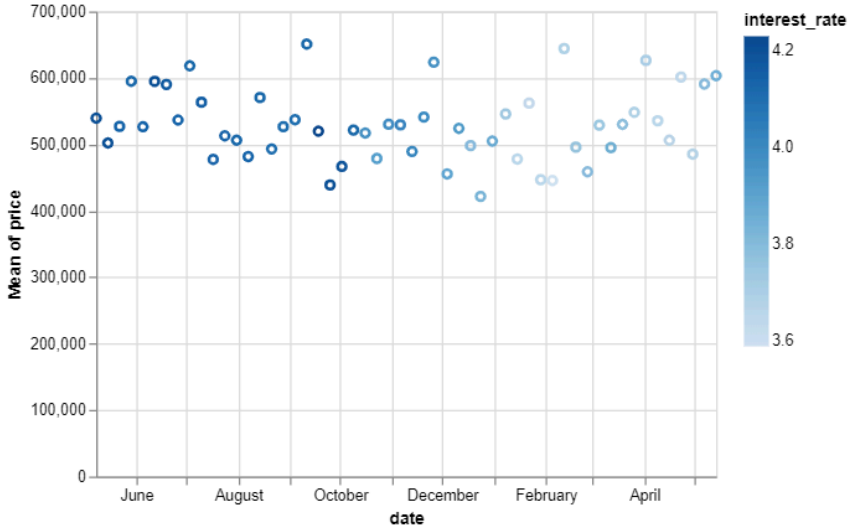
The model would perform better if it was split into two models, one optimized for higher end pricing, and one optimized for lower end pricing, and used only on houses known to be in areas of high or medium-low income.

Gathering More Data

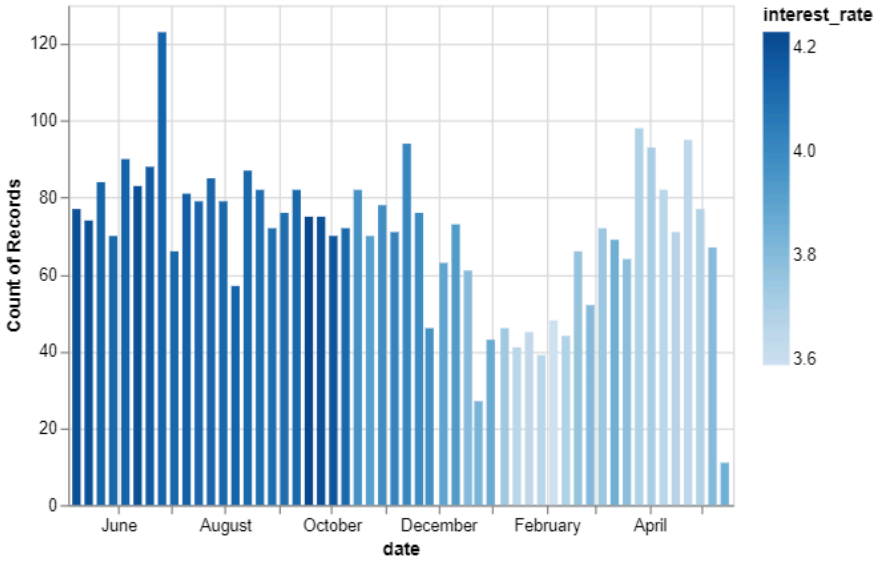
Interest Rates

Interest rates were not included in the data, however, we knew that interest rates have an effect on the selling price of homes. We used the interest rate data from https://fred.stlouisfed.org/series/MORTGAGE30US which contained interest rates for the last 30 years. Since the data that was given was from 2014-2015, those are the years that were used to train our model. The following chart shows how interest rates changed during the time period of the given data. Though there weren't significant changes over the 2 years, interest rates dropped down about 1% from 2014 to the beginning of 2015 and then slowly came back up 0.5%.

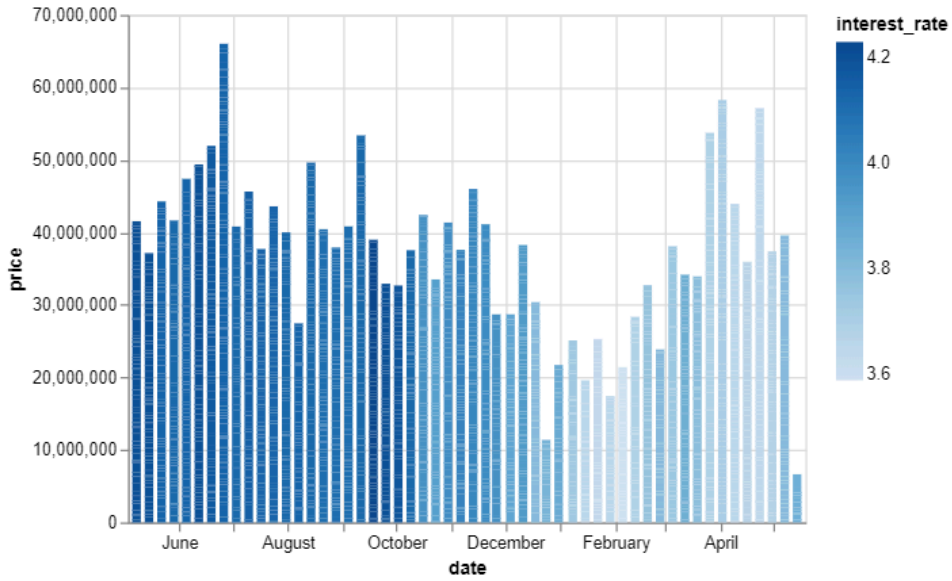
In the following chart, there doesn't appear to be a significant change in average price over the time period as interest rates fall and rise. Just looking at the average though can be deceiving, our model weighted interest rates at a 0.258 in predicting price during this period.



One of the questions we weren't asked to find, but came up when analyzing the correlation between price and interest rates. There is a more significant correlation between the number of homes purchased and interest rates. Which directly impacts this firm to factor in how busy the market will be, and the revenue you could expect. Given more data, we could use machine learning to help predict revenue in the future for the firm.



Interestingly enough, when interest rates were above 4%, there were more homes sold. The sum of the prices follows a similar pattern which we may expect, looking at how the average prices didn't vary too much over time.


70,000,000
60,000,000
50,000,000
40,000,000
30,000,000
20,000,000
10,000,000
0
price
June
August
October
December
February
April
date
interest_rate
4.2
4.0
3.8
3.6