Housing Estimates Project

Executive Summary

Presented by

Jed Chang
Bridger Norman
Celeste Popoca
Luke Russell
Kelin Tang
Ella Yang

I. Summary

As data analysts for Reddic Housing LLC, we received housing data to determine whether we could accurately predict the price of a home based on its various features. Based on these features, we created new columns of data to use on the Gradient Boost method. This method is “a machine learning technique… that gives a prediction model in the form of an ensemble of weak prediction models” [1]. These prediction models, called trees, are created and improved in a sequence using data we already have. We paid special attention to the location, specifically the longitude and latitude, of the housing when we were creating our model. Our calculations lead us to get an R^2 correlation score of 0.90, which is a strong positive correlation. This means our model fits the data’s pattern very well.

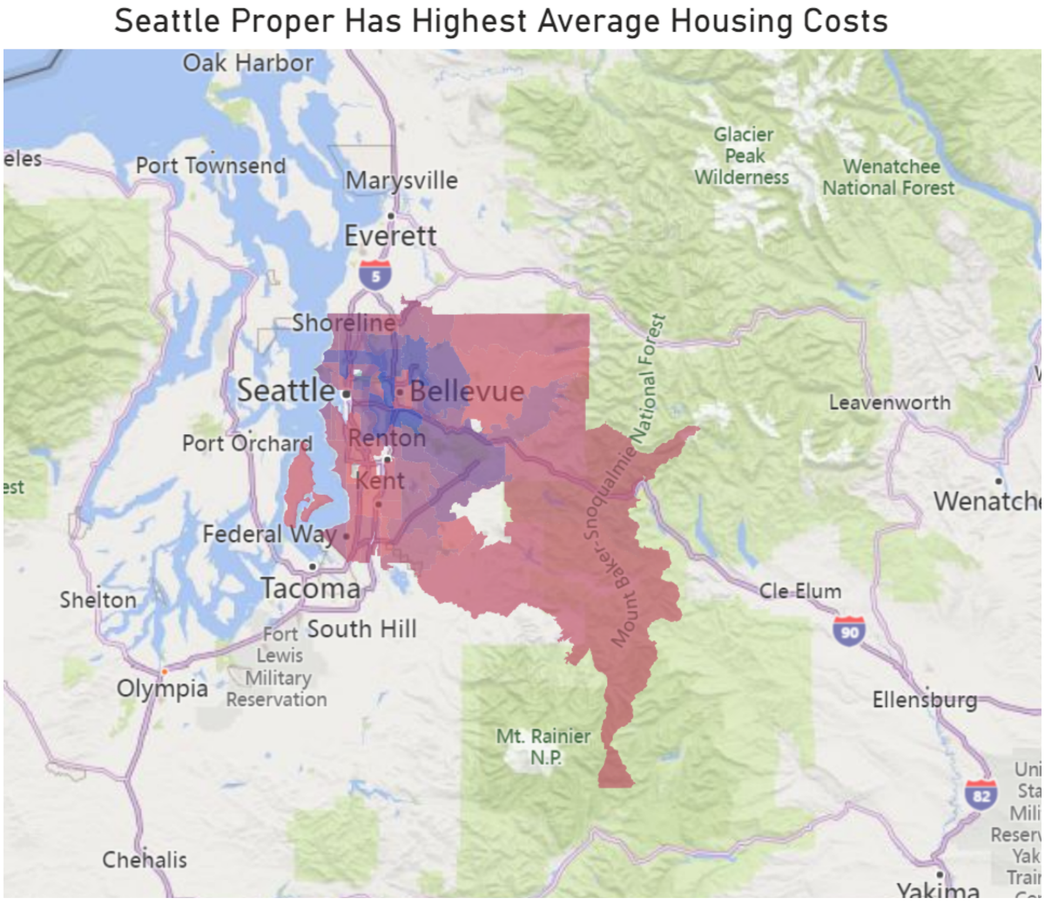

Seattle Proper Has Highest Average Housing Costs

Figure 1

The map above [see Figure 1] shows us how the average housing prices are distributed across different zip codes in the Seattle area. We can identify the approximate center of the highest-priced location based on the colors: blue means a higher average housing price, and red indicates a lower average housing price. In the areas that are blue, we can add a markup to the housing price, whereas in the red areas, we could apply a slight price cut if necessary.

Once we had a model with a good correlation, we tested it against new data to determine how well our model performed against new data. When we ran the test, we got a new R^2 correlation score of 0.90. So, our model performed extremely well once again with the new data. We are confident this model will save the company time and increase sales.

II. Methodology

Process:

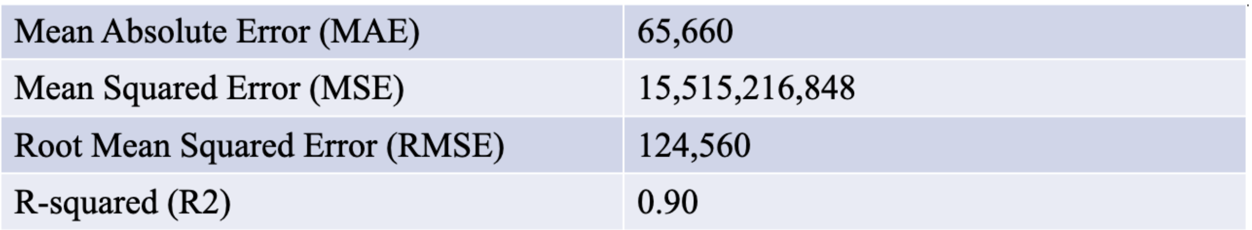
1. Create a vanilla model to know
2. Created graphs and matrix to visualize the data
3. Based on the visualizations, we chose which features would most effectively return the desired output
4. Used feature engineering to create complex features that benefit our model
5. Created multiple gradient-boosted trees and compared their R^2
6. Choose our best model based on performance with a set seed.
7. Tested our model with new data.
8. Determined how well the model predicted the outcome

We began by exploring the data and discovering trends and relationships in the housing data. We then created a vanilla gradient-boosted tree so that we could compare how well our more advanced models would perform. Then we used a correlation heat map and chose our best features. After that, we decided to help our model out by creating 7 new features that provide extra information like rating percentage, housing differences of neighbors, distance from Seattle, and renovation information. We also gathered extra data from the web to account for income by zipcode and we called this feature median age. The new features that influence our pricing estimates are, [‘age_renovation’, ‘median_income’, ‘quality_opinion’, ‘age_of_house’, ‘total_property_sqft_dif_of_neighbors’, ‘was_renovated’]. These are the other original features we used to train our model [‘bedrooms’, ‘bathrooms’, ‘sqft_living’, ‘sqft_lot’, ‘floors’, ‘waterfront’, ‘view’, ‘condition’, ‘grade’, ‘sqft_above’, ‘sqft_basement’, ‘yr_built’, ‘yr_renovated’, ‘zipcode’, ‘lat’, ‘long’, ‘sqft_living15’, ‘sqft_lot15’]

After doing some feature engineering, we created a random forest regression model to compare our XGBoost model against as well as the aforementioned vanilla model. We found that our XGBoost model performed the best against both models meaning that our XGBoost model should provide the best estimates for housing prices. For these reasons, we chose to keep our best XGBoost model and the results of it can be found below.

III. Results, Action Items, and Limitations

The housing prediction model has a high R^2 score of 0.90, which means it has 90% correlation of the variation in housing prices [see Figure 2].



Mean Absolute Error (MAE)	65,660
Mean Squared Error (MSE)	15,515,216,848
Root Mean Squared Error (RMSE)	124,560
R-squared (R2)	0.90

Figure 2

Note:
R-squared(R^2): "R^2 score as a measure of how close the predicted values are to the actual values. It ranges from 0 to 1, where 0 means the model does a poor job of explaining the data, and 1 means the model perfectly predicts the data." [2]
Mean Absolute Error(MAE): It gives us an idea of how much error, on average, we can expect in the model's predictions.
Mean Squared Error(MSE):A way to measure the average squared difference between the predicted values and the actual values. It calculates the average of the squared errors, giving more weight to larger errors.
Root Mean Squared Error(RMSE): A measure of the average difference between the predicted values and the actual values. The RMSE is expressed in the same units as the target variable, making it more easily understandable and comparable to the actual values.

Based on these results, the housing prediction machine learning model demonstrates promising performance for the company. The model is able to provide reasonably accurate estimates of house prices. It can assist the company in providing reliable price estimates to development firms, aiding in decision-making processes, and facilitating effective planning and development strategies.

A. Action Items

Continuously monitor and analyze the housing market trends and dynamics in different zip codes within the Seattle area. Stay informed about changes in property values, market demand, and socio-economic factors that may affect housing prices. This research will help validate the effectiveness of the adjustments and ensure they accurately reflect the market value of properties.

Fine-tune price adjustments based on insights from the average price per square foot across different zip codes.

Consider implementing dynamic pricing models that adapt to changes in the housing market and regularly update price adjustments.

Conduct comprehensive local market analysis to understand specific factors driving property prices in each zip code.

Regularly evaluate the performance and effectiveness of the price adjustments compared to actual market transactions in low-income areas.

Launch a marketing campaign that advertises how our new AI model performs better than Grenic Housing Inc's AI model.

B. Limitations

Choosing which factor to use is important because factors can affect the machine learning model. The housing executive will be based on the location, year build, square footage, and how many bedrooms and bathrooms, etc... We want the prediction to be as close to the actual prices as we can. If the price we predict is too high, that will lead us to lose competitiveness in the market. On the other hand, if the price we predict is lower than the actual value, this will cause us to lose revenue on a potential sale.

IV. Q&A

Looking at the data and our business model, what kind of machine learning problem do you think we're looking at here?

This is a regression problem.

Is there a way we can easily identify properties in low income areas and have the model lower those estimates to protect our insurance customers' interests?

This chart [see Figure 4 below] shows the average price per square foot across different zip codes. It helps identify low-income areas, allowing us to adjust predicted prices accordingly to ensure fair and accurate pricing, protecting our insurance customers' interests.

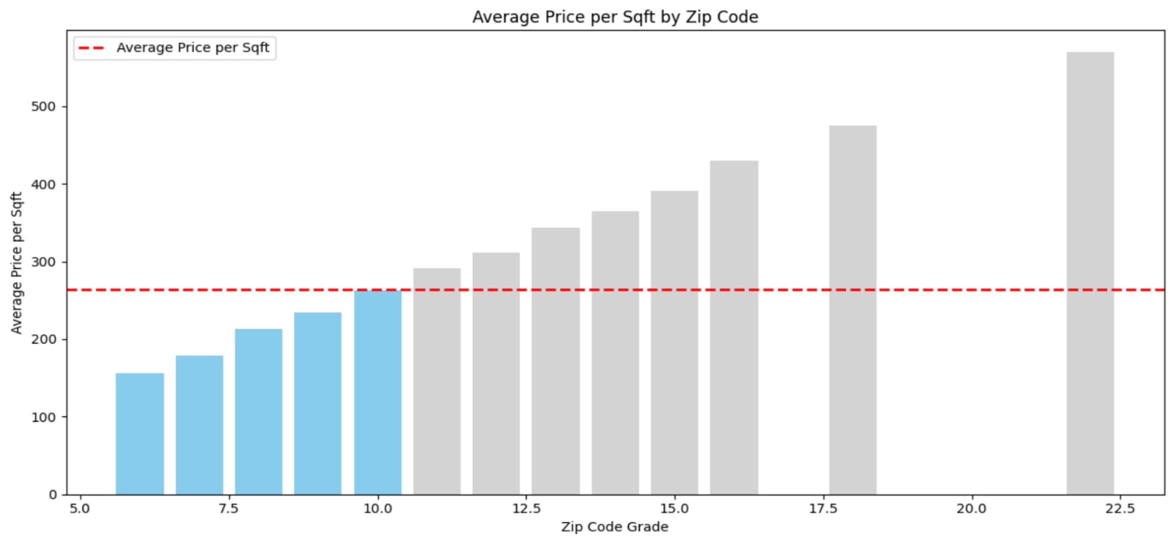


Figure 4

After we added new columns to include adjusted prices [see Figure 5 below] we can see that zip codes with lower grades have a larger difference of original prices. As the zip code grade gets higher, we cannot see much of a difference in price.

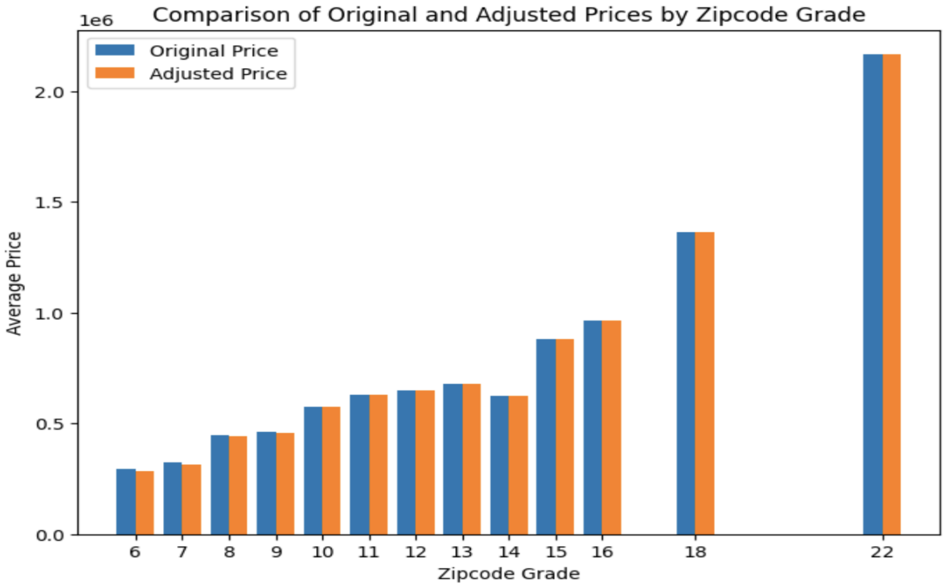


Figure 5

Figure 6 below simply shows a visual representation of where the lower income areas are. The blue area is the highest income, yellow is medium income, and red is lowest income.

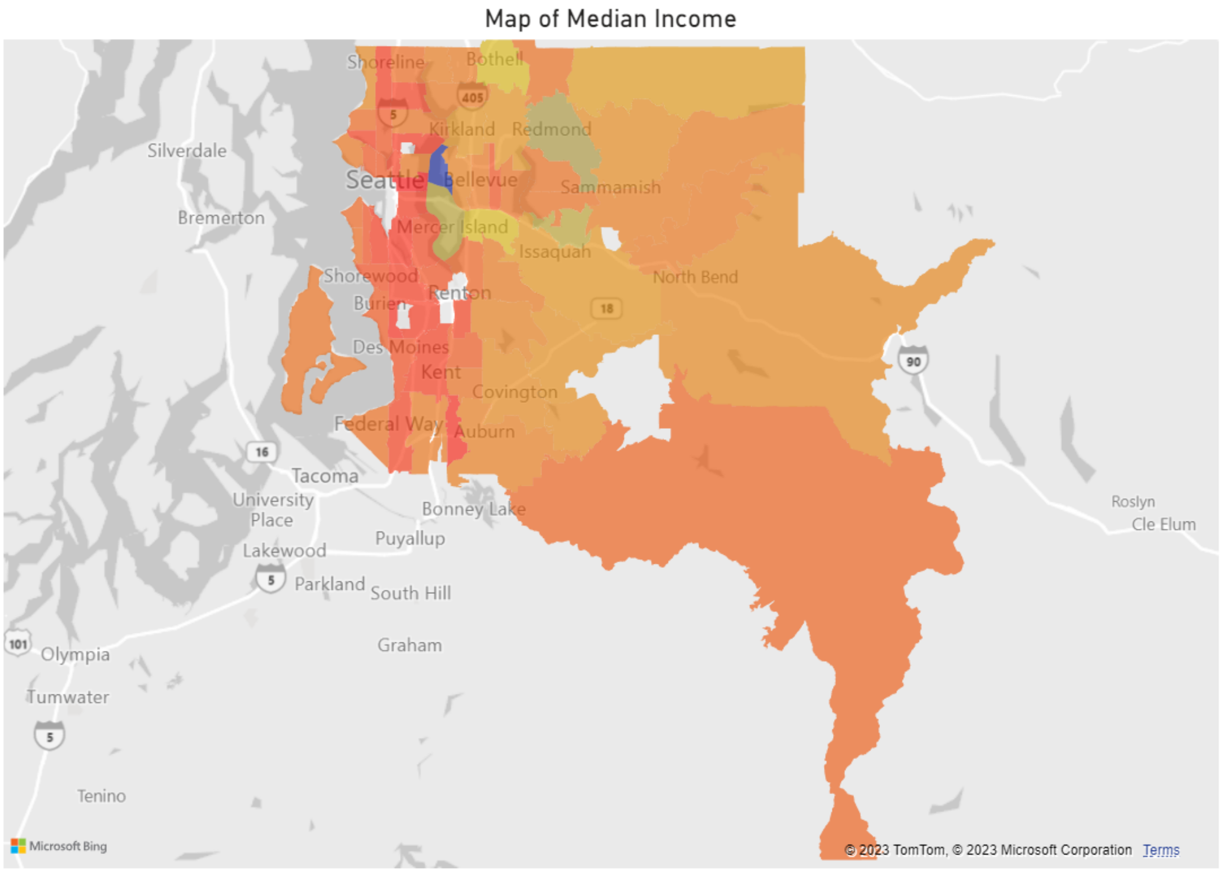


Figure 6

How do we handle square footage?

We created a new column called price_per_sqft_total that is (price/sqft_living).

V. Python Notebooks

Below is Github Gist link to the notebook we used during this case study:

https://colab.research.google.com/drive/1AHvMQdKrqA9BD_5bx1zqzzx11dxjxEDg?usp=sharing

VI. References:

[1] Wikipedia contributors. (2023, May 21). Gradient boosting. In Wikipedia, The Free Encyclopedia. Retrieved 19:56, May 23, 2023, from https://en.wikipedia.org/w/index.php?title=Gradient_boosting&oldid=1156127757

[2] Kelleher, J. D., Mac, N. B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies. MIT Press.