Housing Estimates

Housing Predictions Analysis

Presented by:
Savanna Whittaker
Mahonri Reynolds
Wil Jones
Teghearnan Stevens
Kalina Warner

I. Objective

The goal of this case study was to develop a model that could accurately predict home prices using machine learning techniques. We aimed to explore which features most influence home values and how machine learning could improve prediction performance overall. This summary is intended for Reddic Housing LLC.

II. Key Insights

The findings show that where a house is located, its latitude and longitude, helps predict its price more accurately. These numbers helped the model understand things like how close homes are to good schools, safe areas, or parks, without needing external data.

We also made new features from our data, like how fancy a house is, a luxury score, how old it is, and how good the view is. These helped us understand hidden patterns that affected price.

In the end, we combined two models: a XGBoost and a Random Forest. We adjusted how they worked to make sure they didn't guess too high or too low. Then we blended their predictions together and made small changes to handle prices that were outliers. This gave us much better results overall.

III. Results

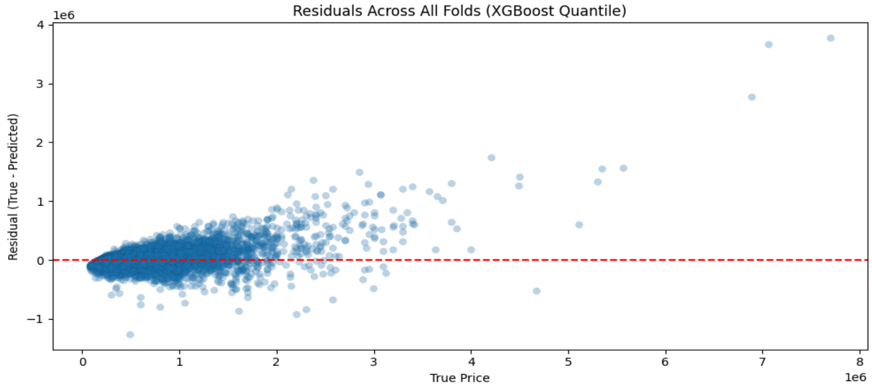
Our model performance on the test set of data, and the final model achieved:

- **Median Error:** $33,842.78
- **Mean Error:** $58,787.90
- **% Median Error:** 7.62%
- **Root Mean Squared Error (RMSE):** $109,329.21
- **R^2 Score:** 0.9108

Overall, our approach predicted home prices with strong accuracy and captured key neighborhood effects without needing external datasets. These results show that our model captures over 91% of the variance in home prices with a Root Mean Squared Error (RMSE) of approximately $109,329.

Figure 1. How Accurate Were Our Predictions?

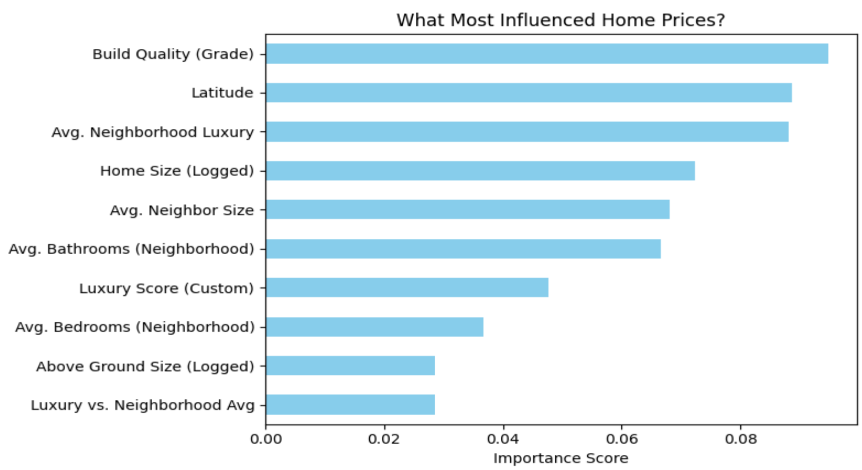
This chart shows how close our predictions were to the real prices. Most are very close, with few minor mistakes.



This shows how close our model came to actual prices. Dots near the red line mean better predictions.

Figure 2. What Most Influenced Home Prices?

This chart shows which things affected house prices the most.



This shows the most important features our model looked at when predicting prices. Things like build-quality and location mattered most.

IV. Limitations

While the model performed well overall, there were limitations. For example, some important external data like income levels or exact school ratings weren't available. Also, rare luxury homes were harder to predict because they behave differently than most homes.

V. Python Notebooks

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

[View Our Notebook on Google Colab]

VI. Discussion Responses

- This is a regression problem, as we're predicting a continuous value—housing prices
- RMSE is useful because it communicates errors in dollar terms and emphasizes larger mistakes and will be able to show more confidence to the stakeholders.
- To ensure fairness, especially in low-income areas, we should integrate external data like income levels, poverty rates, and homeownership statistics as features. However, we should show these features as part of our data, but not manipulate the rates.
- XGBoost and random forests do not require feature scaling; this would be easier to use so that standardization and normalization do not have to be considered for our model.
- XGBoost is robust to unscaled features and can handle skewed data effectively. Applying transformations like log scaling can further improve model performance and pattern recognition.