Housing Estimates, Project

Presented by

David Orizu

Introduction

Reddic Housing LLC specializes in providing house price estimates to development firms. To maintain a competitive advantage, they are embracing innovative ways in machine learning to enhance their price estimation capabilities for homes. This initiative aims to leverage modern data science techniques to provide more accurate and reliable property valuations, ensuring Reddic Housing remains a leader in the real estate market.

Our primary objective is to develop a machine learning model that not only predicts housing prices with high accuracy but also provides quantifiable metrics of confidence. This model will address the key concerns including:

- Ensuring that our customers can place a high degree of confidence in our price estimates.
- Demonstrating the transparency and interpretability of our machine learning model, comparable to traditional Excel-based analyses.
- Identifying additional factors affecting property prices and incorporating them into our model to enhance its predictive power.

This report outlines the development process of our machine learning model, the insights gained from our data analysis, and the steps taken to ensure the reliability and interpretability of our predictions.

Approach

Our team got straight into using the XGBRegressor for our data, which is a machine-learning model similar to a random forest but which learns from experience as more data is fed into it. We first looked at the hyperparameters within our Regressor and tested it with many different setups. We would keep track of which parameter gave us the best results and adjust our model accordingly. From this, we were able to see at what point each parameter was at its best.

Figure 1

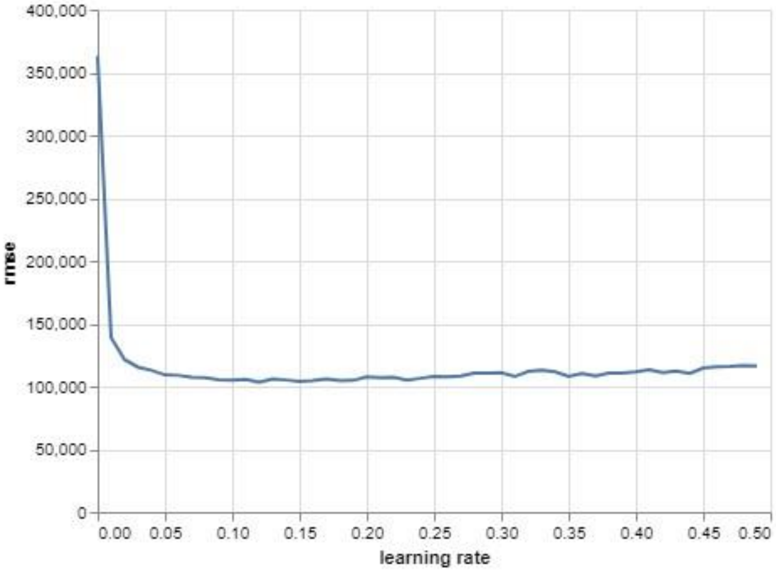


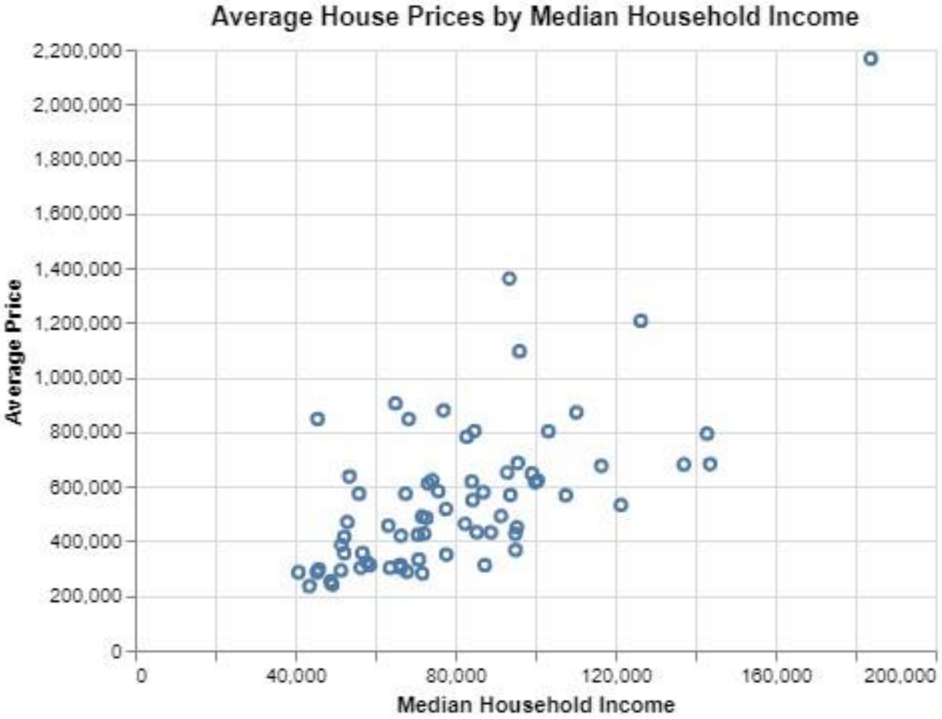
Figure 1 was testing the parameter “learning rate”. The learning rate helps us control how conservative our model will be. You can see how it gets significantly better from 0 to 0.05. After that, it starts leveling out and rising back up. This chart helped us find that the best learning rate was 0.12. We followed this same process for many other parameters.

We then took a closer look at our data and made some changes. We removed a couple of outliers that were skewing our results. We also manipulated our data a little by turning our “yr_renovated” column into a “since_renovation” column by subtracting the year renovated from the current year. If the home had never been renovated before, we took the year it was built and subtracted it from the current year.

Lastly, we added a couple of different features. We added median house income and property crime rates by zip code. This way our model would be better at identifying properties in low-income areas and not overestimating their value.

Figure 2 shows the relationship between the price of a house and one of our new features, Median Household income.

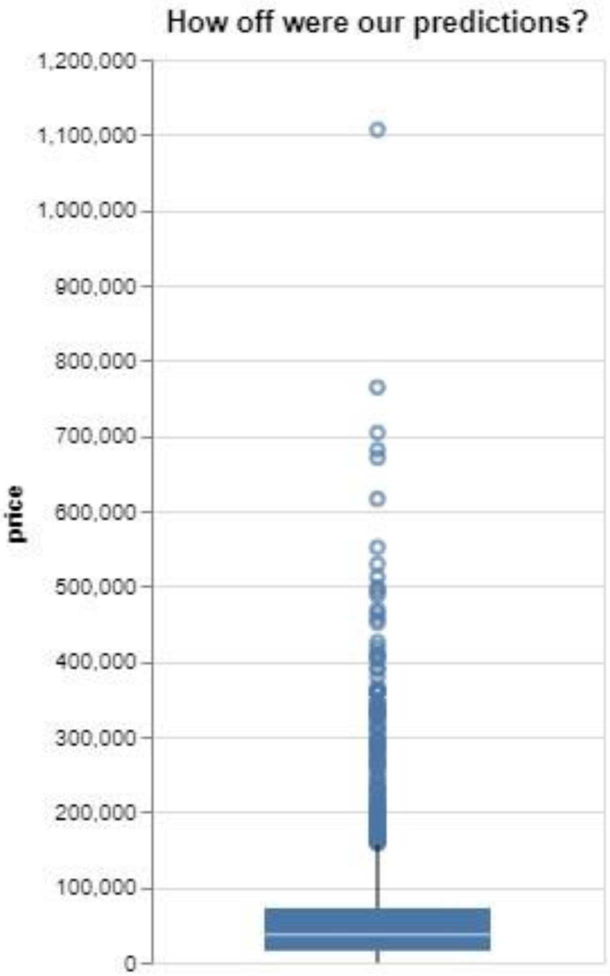
Figure 2



From Figure 2, we see that as the Median household income of a particular region increases, the average price of the houses would increase. This variable shows a solid relationship with the housing price, which would be a good feature for our model.

Model performance

Figure 3



Min: 7

Q1: 15,000

Median: 37,000

Q3: 72,000

Max: 1,100,000

RMSE: 101,443

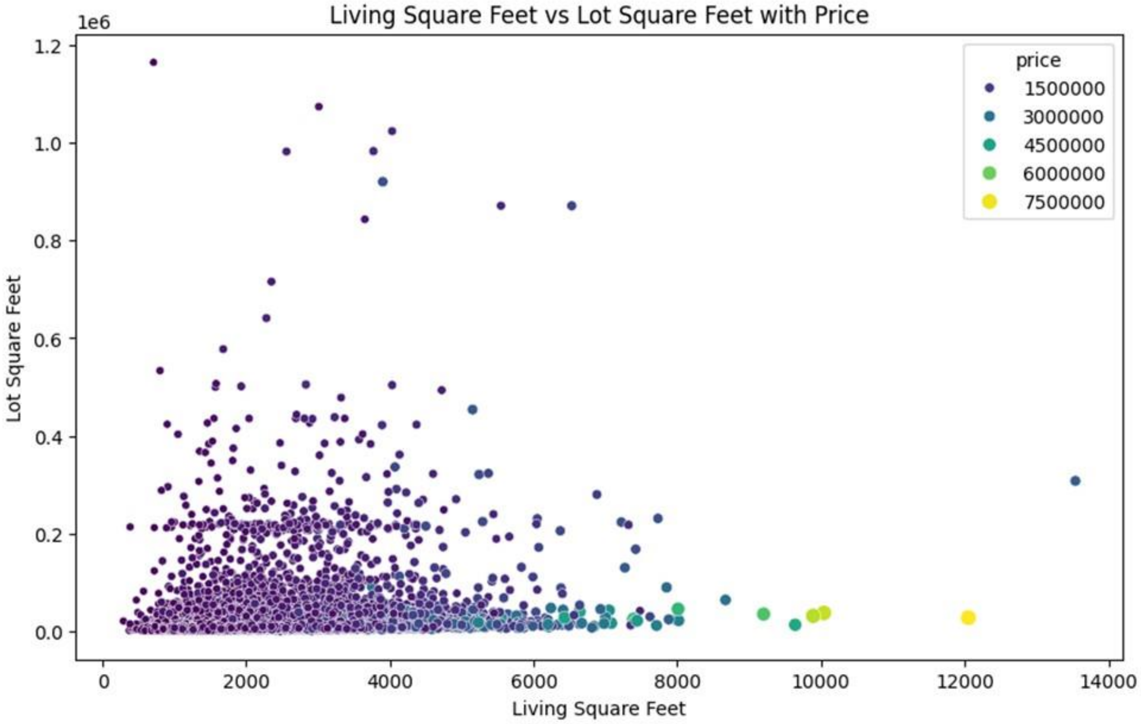
R Squared: 0.922

Figure 3 is a boxplot of our absolute errors with our measurements to the right. RMSE (Root Mean Squared Error) and R-squared are the main measurements we're focusing on, but we thought it would be nice to include the spread of our errors. RMSE is how off we are on average, but it considers outliers more than just looking at the absolute errors. The R Squared is a scale from 0 to 1 that shows how much correlation there is between our data and what we're predicting.

Identifying interesting trends

Our team undertook a thorough analysis of the housing data to identify trends and improve the accuracy of our price prediction model. Here are the key findings and actions we took:

Figure 4



1. Lot Size vs. Living Size:

- Finding: We discovered that the lot size (the land area) of a property is not a reliable predictor of its price.
- Action: We focused on living size (the interior living space), which showed a strong correlation with price. Bigger living areas tend to mean higher prices.

Figure 5

2. Bedrooms and Bathrooms:

- Finding: Generally, more bedrooms and bathrooms correlate with higher house prices. However, we noticed some outliers that distorted this trend.
- Action: For instance, we found a house with an unrealistic 33 bedrooms in figure 5. We removed this outlier to make our data more accurate and reflective of typical properties.

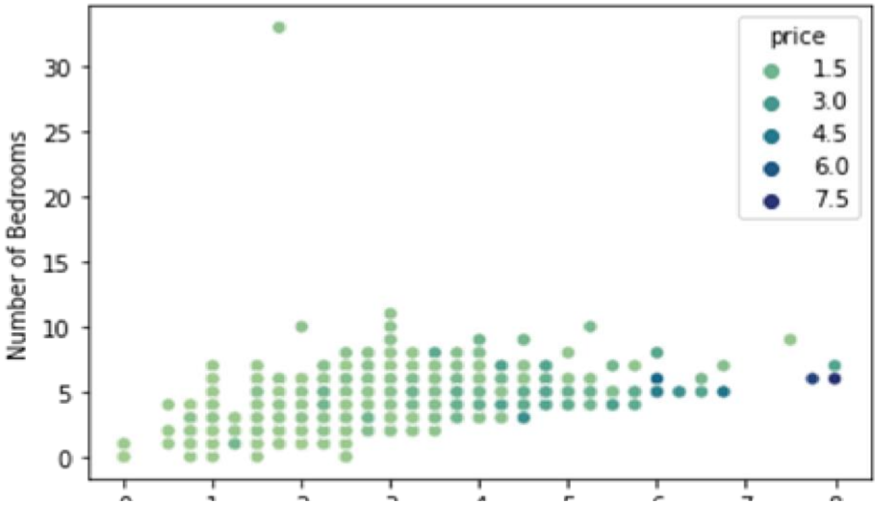
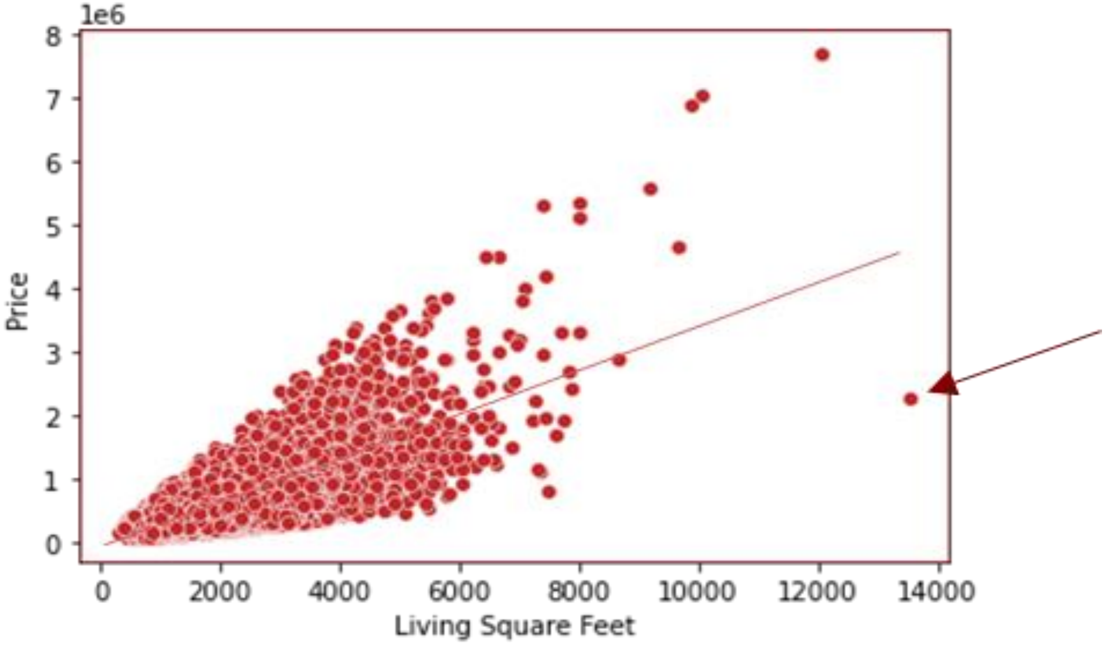


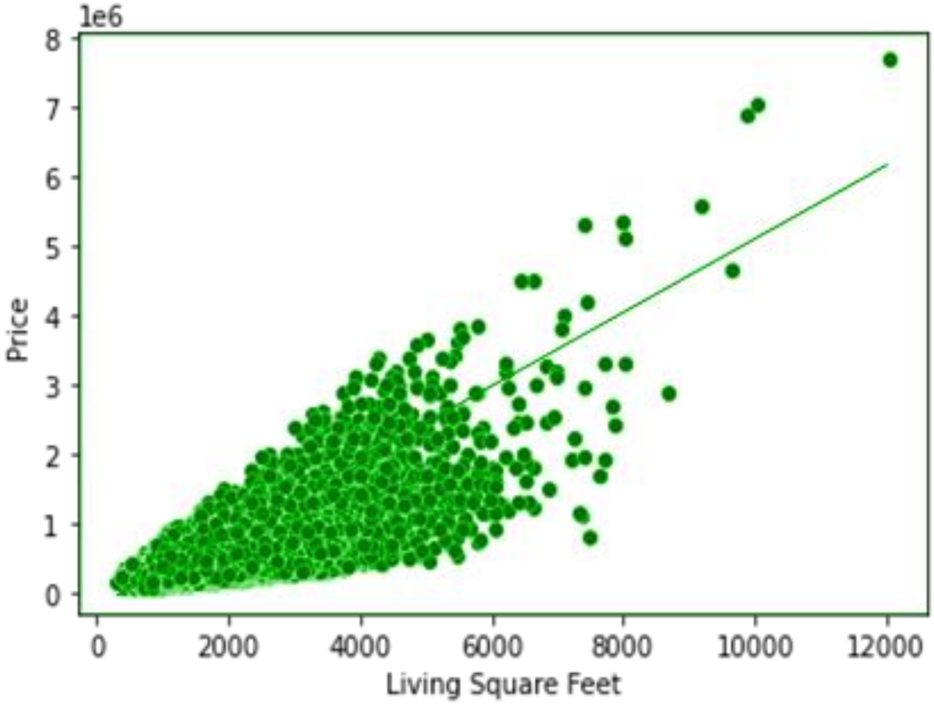
Figure 6



3. Outlier in Living Square Footage:

- Finding: We also identified a house with a large living area but an unusually low price in figure 6.
- Action: Removing this outlier helped clarify the strong relationship between living space and price, making our model's predictions more accurate.

Figure 7

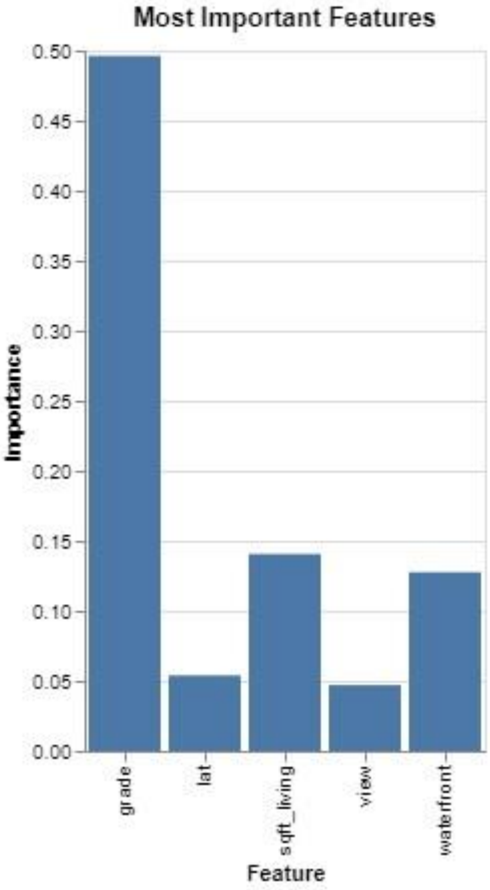


4. Living Square Footage without Outlier:

- Focus on What Matters: We learned how much land a house has (lot size) doesn't affect its price much. Instead, the size of the house itself (living space) is a much better indicator of price.
- Cleaning Up the Data: We found some unusual data points, like a house with 33 bedrooms, which don't reflect typical homes. By removing these odd cases, we made our data cleaner and our predictions better.
- Improving Accuracy: By focusing on important factors like the size of the living space and removing outliers, our model can now predict house prices more accurately.

These steps have significantly improved our model's ability to predict house prices, giving us more reliable and actionable insights.

Figure 8

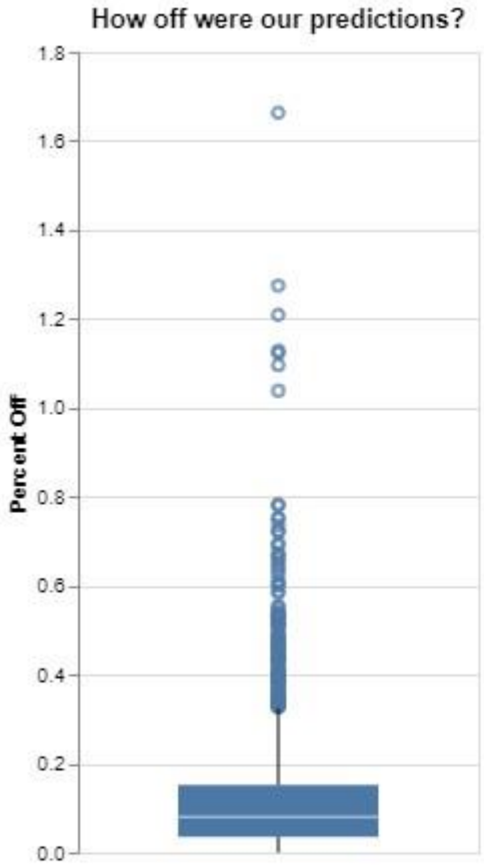


5. Feature Importance:

Figure 8 shows which 5 features played the biggest role in our predictions. You can see grade was the most important, followed by square foot living, and waterfront. By understanding which features are most important in our predictions we can better understand what data is most important to collect for future homes.

Showing the benefits of technology for the company

Figure 9



Although the Root Mean Squared Error was what we measured, Figure 9 gives us an even better understanding of how accurate our model will be. This chart illustrates that 75% of our predictions won't be more than 15% off and have a median of 8% off.

Conclusion

Overall, our efforts successfully demonstrate the value of machine learning in generating accurate and understandable house price estimates.

- Our XGBoost model achieved an R-squared value of 0.922 and an RMSE of 101,443, indicating strong predictive accuracy, with 75% of predictions falling within 8% of the actual price.
- By analyzing feature importance, we identified key factors influencing house prices (e.g., grade, living square footage, waterfront location), guiding future data collection efforts.
- Including diverse data sources, such as median household income and property crime rates by zip code, improved model accuracy. Exploring additional external data sources in the future could further enhance predictive power.

Plans moving forward.

- Following the release of this model, further fine tuning and adding more data will greatly improve its performance on future data.
- Continuously monitor and refine the model with new data to maintain its accuracy and effectiveness over time.
- The success of this model in the Seattle area suggests its potential for application in other geographic zones with similar housing markets. Expanding testing to these new areas would enhance the model's generalizability and broaden the reach of Reddic Housing's services.

By combining these advanced algorithms with human expertise in data cleaning, analysis and learning, Reddic Housing LLC can stay ahead of the curve in the competitive real estate market.

Google collab Notebooks

https://colab.research.google.com/drive/1ub_IvCXRruIhFat22L37sm0T469-mFkh?usp=sharing