Module 3: Executive Summary

Marketing Campaign Analysis

Presented by:

Joaquin Solis

Scott Townsend

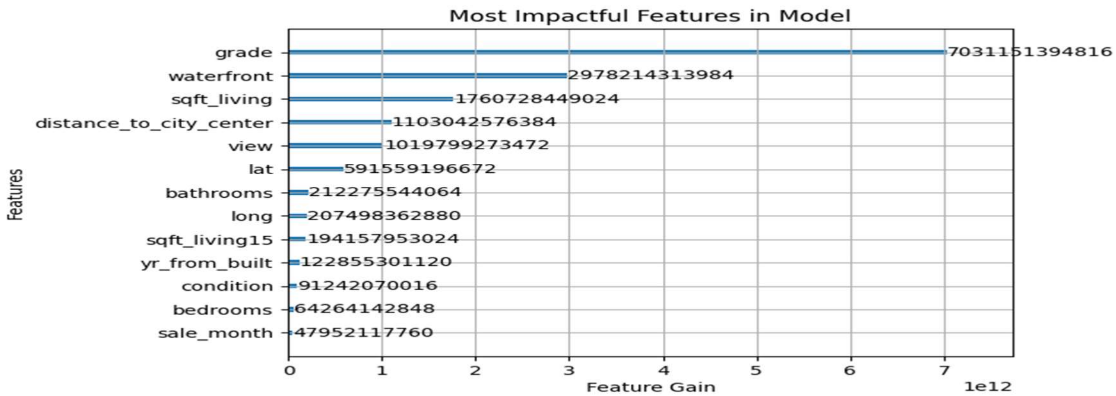
Zach Holcomb

Isaac Weyland

I. The Data

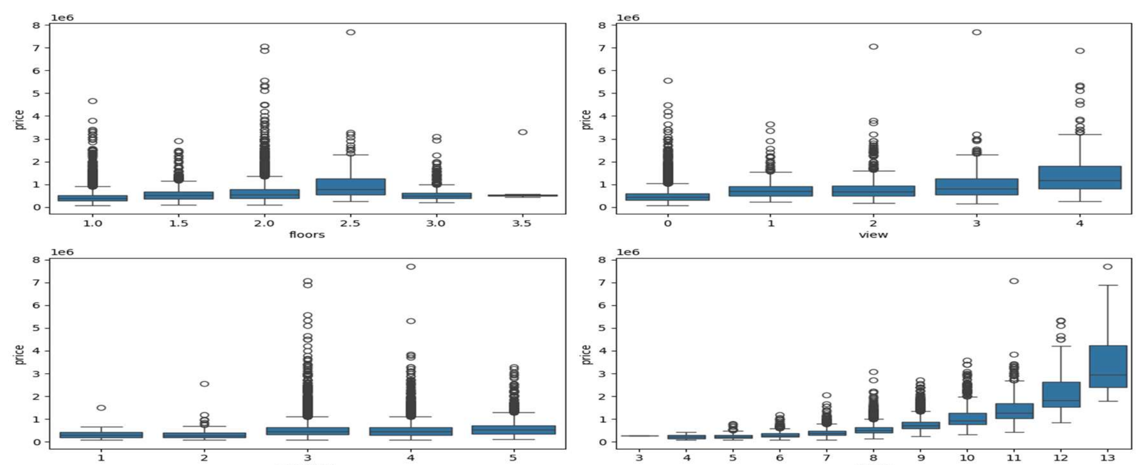
As we reviewed the clients' meeting notes, we noticed that there were some questions raised about the data. Before we dive into the details of the model we trained and developed, we will explore some the insights, features, and responses to questions that our model found:

Impactful Features on Model



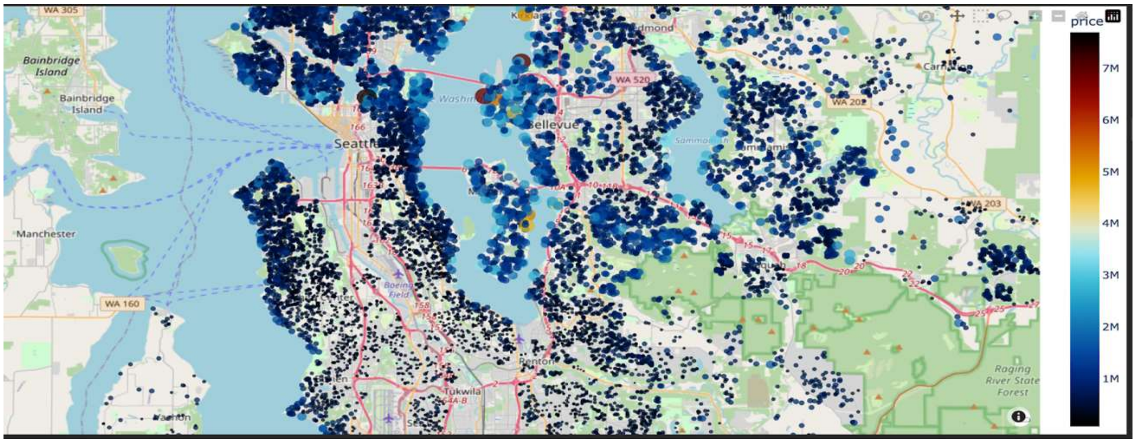
The question was raised, on which property types are weighing most heavily in the house prices predicted by the model. The above bar chart displays all the most impactful features on predicting home prices.

Feature Influence on Housing Prices



These above boxplots compare house prices with different features. We can see a correlation - that homes with more floors, better views, in better condition and higher grade tend to have higher prices.

Seattle Homes Heat Map



The above heat map demonstrates a few things. First, there appears to be clustering of higher-priced homes in certain neighborhoods, particularly near the waterfronts. We also see some higher priced homes appear more inland, which shows a variability in pricing.

The question of: "If there are additional factors about these areas that might be affecting prices, which we aren't taking into account" – While there could be additional factors that we aren't considering, there were none that we were able to find while analyzing the data.

II. ML Model

Our model was trained and developed using a Seattle housing data set. We had to first clean the data before we could begin analyzing and training our model. This process included converting date variables, removing outliers, and finding meaningful data to add.

Training and Testing Strategy: We decided to split the data, using 90% of the data for training the model and the remaining 10% for testing. To improve the model's performance, we used a boosted regression forest model to predict the exact price of the homes. After creating the model, we tuned the features and parameters to improve it focusing on RMSE.

Model Performance: We used the XGBoost Regressor, which works well for predicting continuous values like housing prices. We measured its performance using two metrics - Root Mean Squared Error (RMSE) which will show us the average of our predictions compared to the actual prices. Additionally, we will use R-squared (R^2) which will display how well the model fits the data, or rather how accurate it is. **We achieved an RMSE of 104,188.3 and R-squared of 0.91** which is quantifiable evidence of the accuracy of our model.

Challenges: Some challenges with this case study included finding meaningful data to add to the model and deciding which variables to use. When we added data to our model, we used the error metrics as a parameter to see if the performance improved. We repeated this process several times until we found the optimal variables.

Best Use Case: This model can speed up real estate investing by quickly estimating prices for newly listed homes. Sellers and buyers can use it to check prices more accurately. Real estate agents can also use it to spot homes likely to sell at higher prices, especially those in good locations.

III. Python Notebooks

Notebook

IV. Discussion Responses

1. What type of machine learning problem is this?

With this data set, the machine learning problem would be *regression* as housing prices are a continuous variable. The goal of our machine learning model is to predict housing prices in a certain area with variables like square footage, location, and the number of rooms, and a regression model will help us analyze how different variables mentioned above affect the price of a house.

2.How should we evaluate our model's performance?

To best validate the model's performance, our team recommends using *Root Mean Squared Error (RMSE)*. RMSE will give us a direct measure of how well the model's predictions align with the actual values, offering clear insights into the accuracy of the model in a way that will be simple to understand.

3.Can we adjust predictions for properties in low-income areas to protect our customers?

No, adjusting model predictions based on income levels would violate *federal laws and ethical standards*. The *Fair Housing Act* prohibits discrimination based on income, race, and other protected characteristics.

4.How should we handle the varying scales of features, like square footage, in a gradient-boosted tree model?

While gradient-boosted trees don't strictly require scaling, it can be beneficial for interpretability. For features like square footage, you can scale the values by dividing by a factor of 10 or 100 to bring them closer to the ranges of other variables, ensuring smoother model performance without large discrepancies in feature scales.