Executive Summary

Marketing Campaign Machine Learning Analysis Report

## Presented By:

Becca Roeth

Ethan Houser

Kyle Bell

Kyle Parks

Makayla Whipple

Chantal Borchardt

# I. Introduction Summary

In the highly competitive real estate market, accurately pricing homes is critical for maximizing revenue and meeting market demand. Our team has utilized advanced machine learning techniques to develop a **Gradient Boosted Regression model**, designed specifically to predict housing prices in King County with [insert accuracy statement here]. By leveraging the rich dataset available for King County, this model offers a sophisticated and reliable method for pricing homes.

Gradient Boosted Regression is a powerful machine learning algorithm known for its robustness and accuracy in predictive modeling. It builds an ensemble of decision trees in a sequential manner, where each subsequent tree aims to correct the errors of the previous ones. This iterative approach allows the model to capture complex patterns and interactions in the data, resulting in highly accurate predictions. Here’s why this model is effective for determining housing prices:

1. **Feature Importance**: Features are data points that the model is trained on, such as square footage, floors, bathrooms, etc.. The model inherently evaluates the importance of each feature in predicting housing prices, allowing us to focus on the factors that have the greatest influence. This insight is crucial for making informed pricing decisions.
2. **Robustness to Outliers**: The algorithm is less sensitive to outliers compared to other regression models, ensuring that the predictions remain reliable even in the presence of anomalous data points.
3. **Flexibility and Customization**: Gradient Boosted Regression can be fine-tuned through various hyperparameters, allowing us to optimize the model specifically for the King County housing market. This customization enhances the model’s predictive power and adaptability to market changes.

# II. Addressing Questions

1.1 Problem Type

This is a **Regression** problem (predicting numeric value) and so the best Learning Module would be a **Supervised Learning Algorithm** (which can predict future data points from labeled pre existing data points).

Because the XGBoost model is made up of **Decision Trees** (series of different choices made from data, Ex: if I have a 75% in the class, I am not failing but within that I do not have higher than 90% which is not an A), the data does not need to be scaled, (unlike with other models we could use to solve a *regression* problem like this one). That means we don’t have to worry about features with different ranges like square footage.

1.2 Model Confidence

After looking at the data we were given it’s evident that we need to use a **Gradient Boosting Regression Model** (a gradual, careful model made up of multiple decision trees that are each created on past inadequacies). Overall, when it came to predicting house prices the confidence metric we chose to use was the **R^2** (a scoring method with results between [0,1] with 1 being the highest and best performing score).

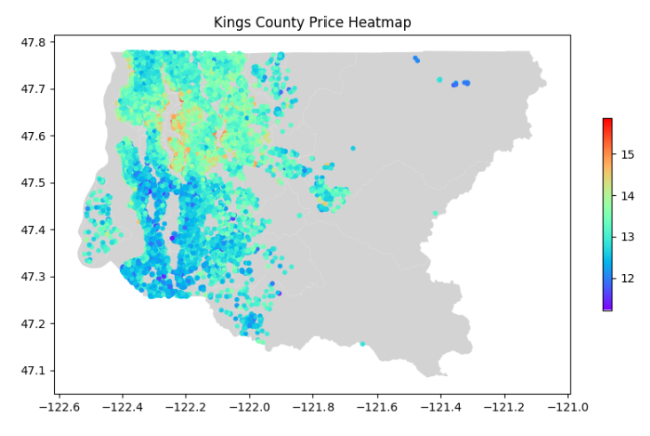
With the R^2, we are able to show how well our model explains variability of house prices without needing the context of the dataset to know what is “good” or not (for example, the Mean Absolute Error finds the average difference between an estimate and the actual value. However, it requires intimate contextual knowledge of the dataset to know if the score is “good”). While we will use a variety of score metrics, we compared which model was best with R^2.

1.3 Insurance Question

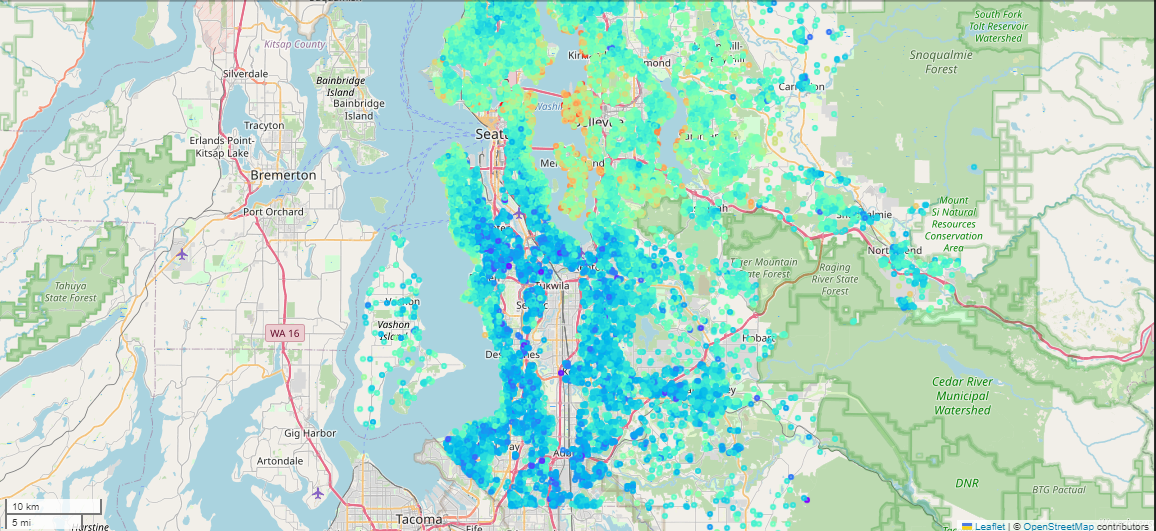
Regarding the request to lower estimates to protect our insurance customers’ interests: **we have chosen to exclude this information**. We can estimate the current price based on the location/neighborhood of the house, but we legally and ethically cannot *lower* the price of the house based on if it’s an ‘unsavory’ neighborhood. *Further explained in limitations.*

# 

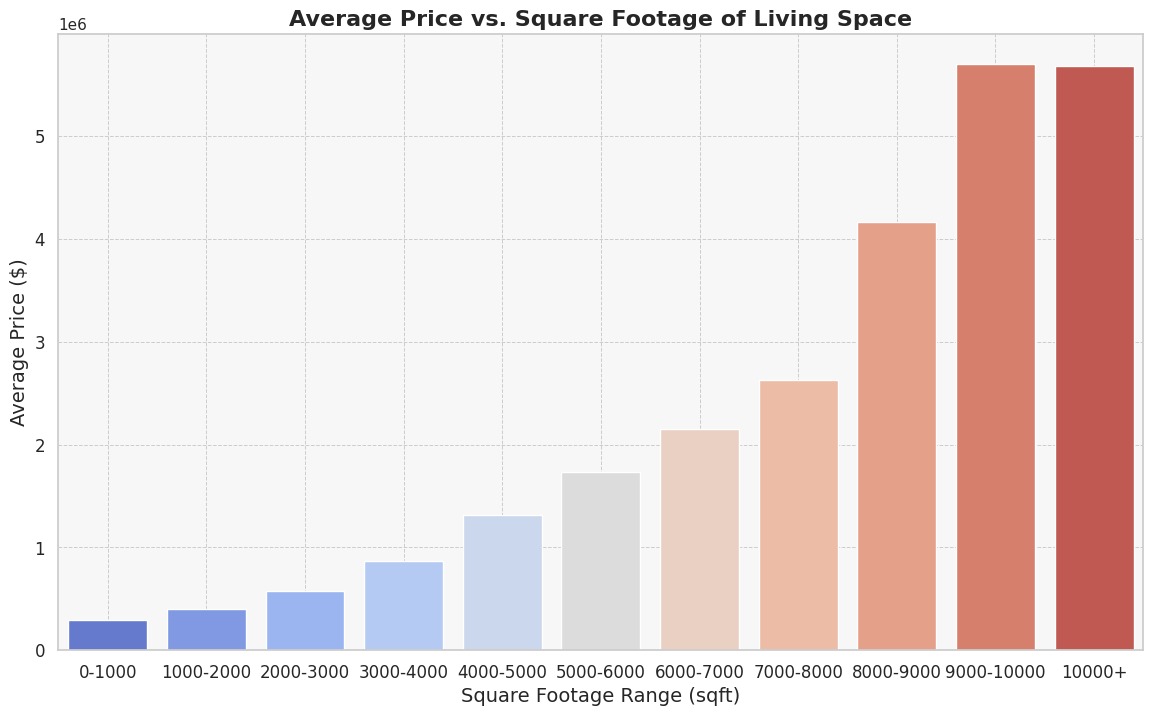
# III. Data Exploration

*Graph 1*

*House prices from King’s County, Washington [Seattle Area], from May 2014 - May 2015 (axes are Latitude, Longitude). Shows the location of high/low priced houses and that south is lower priced while north is high priced.*

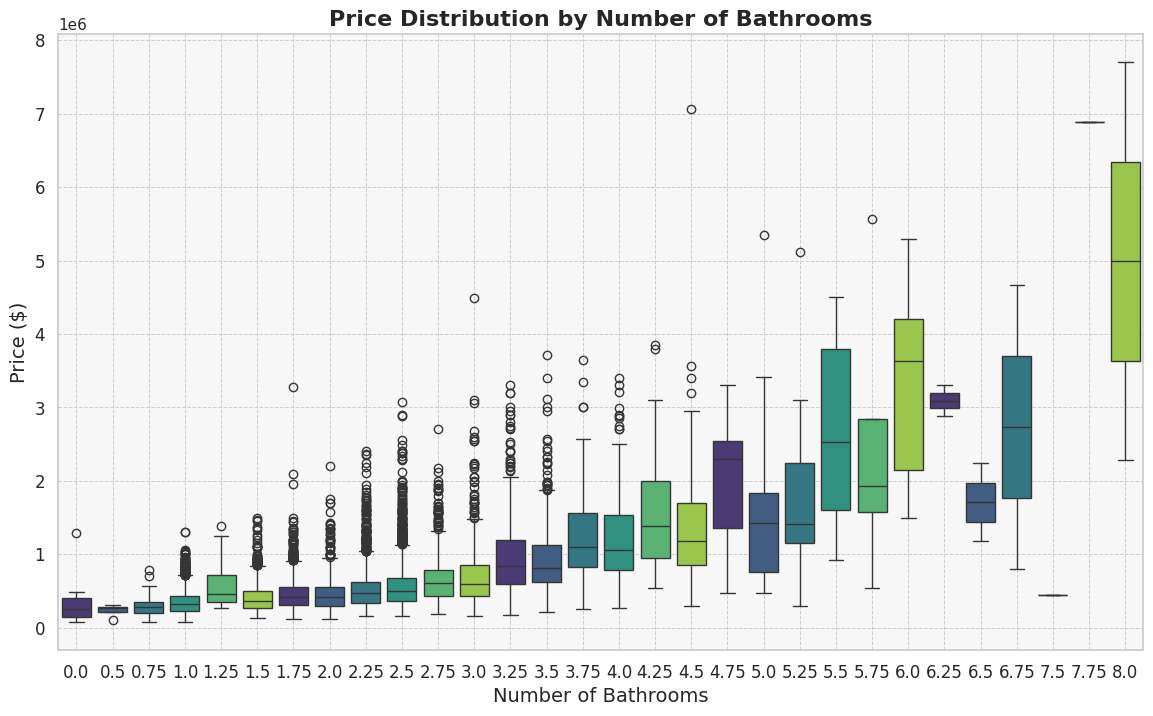
*Graph 2*

*House prices from King’s County, Washington [Seattle Area] on a geographical map, from May 2014 - May 2015 (axes are Latitude, Longitude). As the location of housing nears the greater Seattle area, the pricing of the house increases. This also shows that houses near a waterfront have a higher value.*



*Graph 3 (Price in millions)*

*As the square footage increases, there is a major increase in housing value.*



*Graph 4 (price in millions)*

*The number of bathrooms has a small influence on the price when there are 3 or less bathrooms, but above that the price of houses increases dramatically.*

## Issues with the Data

* The data shows a **higher concentration of houses priced below $1 million** compared to those priced above this threshold. This imbalance could *negatively impact the accuracy of our AI model when predicting prices for higher-valued homes* because the model is trained on a disproportionate amount of lower-priced data.
* Certain features in the dataset contained **outliers** that *significantly skewed the data*, such as a house reportedly having 33 bedrooms. *These anomalies needed to be excluded to enhance the model's accuracy*. By removing these extreme outliers, we can ensure that the model is trained on more representative data, thereby improving its predictive performance.

# IV. Machine Learning Model

## Feature Importance

Feature importance refers to the process of *ranking features based on their impact on the performance* of a machine learning model. This is important for avoiding redundancy and overfitting while also learning generalizations so we can apply this model to future datasets.

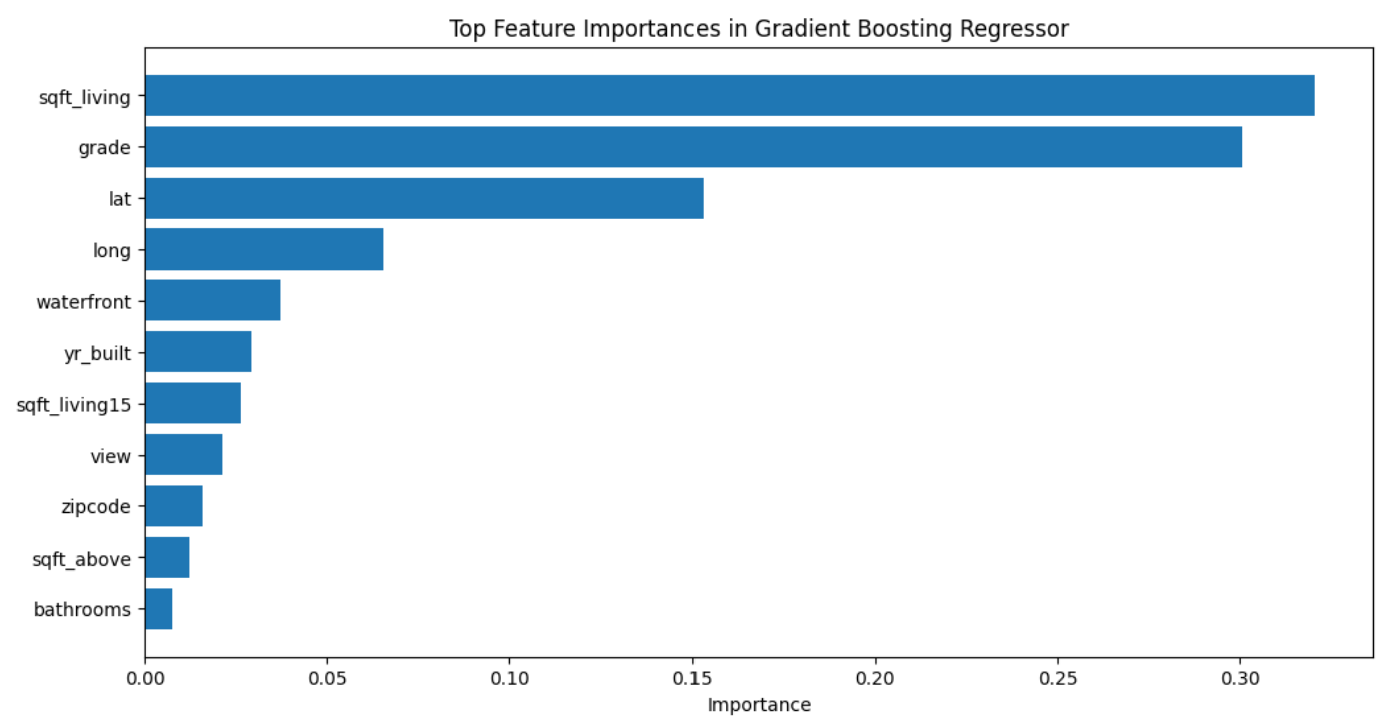
**The chosen features are the factors that are proven to help contribute to the model’s accuracy**. That is why they are considered important. Those deemed unimportant did not help the accuracy of the model. In some cases, like the bedroomscolumn that tracks the number of bedrooms per house, removing a feature improves the quality of the model. Not all information is useful when it comes to machine learning.

*Important* ***(Graph 5)***

* SQFT of living space
* grade of the house
* LOCATION
  + Latitude - The further north you go in King’s County, the closer you get to downtown Seattle and to better-off neighborhoods
  + Longitude - Less important, but prices of homes above a certain latitude get cheaper when further from downtown
  + zipcode - a more generalized representation of the location of the house
* Whether the apartment overlooks the waterfront
* Sqft of 15 Nearest Neighbors
* View - An index from 0 to 4 of how good the view of the property was
* SQFT of housing space above ground level
* Number of Bathrooms (**Graph 4**)

*Unimportant*

* Number of bedrooms
* Square footage of land on the lot
* Number of floors - (accounted for in sqft)
* Square footage of basement - (flood/mold risk per Seattle climate)



Graph 5

*The feature importance graph was able to help give us a visualization of* ***what values contribute most to price.*** *This allowed us to test these features and understand their impact on the house price. Some key features we chose to use were bathrooms, sqft\_living, waterfront, view, condition, grade, sqft\_above, yr\_built, yr\_renovated, zipcode, lat, long, sqft\_living15, sqft\_lot15, and month.*

## Feature Engineering

**month**

The month the house was sold in, derived from the date feature.

**Average\_Household\_Income**

This new column takes the average household income for each zip code area. This came from incomebyzipcode.com and it included average household prices for the year 2022.

**prices\_in\_zip**

This includes a list of the prices of other homes included in the zipcode. This came from the price feature from multiple house instances.

# V. Results, Action Items, and Limitations

## Results

| **R^2** | 0.88 |
| --- | --- |
| **Root Mean Squared Error (RMSE)** | $134,091.81 |

## Action Items & Conclusions

* **Seattle area housing prices tend to stay relatively stable**. The market may have its ups and downs, but it quickly corrects back to normal.
  + However, this data was from a two year period and could be biased so gathering more data from a wider time-line would be advised
* Deliberately lowering the predicted prices in certain neighborhoods would undermine the integrity of the model, and it is also illegal on the federal level.
* **Time of year greatly impacts housing prices.** In the winter [Nov-March], there was lower demand for homes further from the city center. People prefer to move in the warmer months.
* **The number of bathrooms contributes to house value** much *more* than the number of bedrooms.

## Limitations

1.3 Insurance Question *(continued)*

We were able to find and utilize the average salary per zip code and use that as a way to determine neighborhood wealth. We used this information to make a more accurate prediction using our AI model. However, *we cannot use this information to explicitly lower the pricing estimate on a house due to the Fair Housing Act*.

Under the **Fair Housing Act**, lowering predicted prices for homes in specific neighborhoods can be seen as discriminatory, intentional or not, if these neighborhoods are predominantly composed of protected classes. *The FHA prohibits practices that result in disparate treatment or have a disparate impact on protected classes*. If your pricing model adjusts values based on neighborhood characteristics, it may inadvertently result in discriminatory practices against these groups.

* However, in any dataset involving a system of class disparity, there are ever-present implicit biases in the data that disfavor “unsavory neighborhoods”.

# VI. Python Notebooks

Master: <https://colab.research.google.com/drive/10qgozXsVKyBqW-kYx9-YQ8GeVPGcLtMn?usp=sharing>

Fancy Maps:

<https://drive.google.com/file/d/1SnX_JtUCumOm54qfD1H1BQ5FW3GeV2Iu/view?usp=sharing>