

Housing Price Case Study

presented by

Ezekial Curran

Isaac Bojorquez

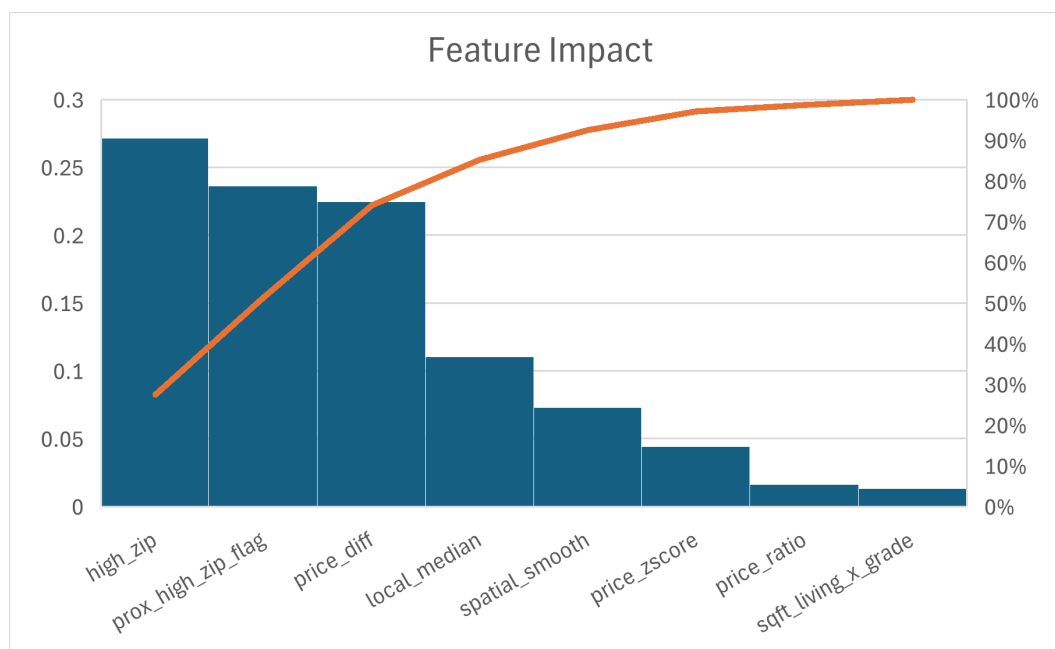
Tonatiuh De Leon Carvajal

I. Perspiration to prediction

This study evaluated the provided housing information to predict and anticipate housing prices and the key attributes influencing housing prices. The dataset has around 20,000 rows of data with each representing a housing and what it sold for. Due to the size of the dataset we can use the different characteristics of the housing options to better predict a price.

II. Greater prediction accuracy

Some of the top characteristics that influenced the price of the housing were grade, location, and square footage of living space. Additional data was included from the census related to the relevant zip codes for the relevant years and it was found that education had a significant influence as well. The introduction of education data such as bachelor's degrees and high schools in a zip code increased the accuracy. The grade follows public opinions and what people value as well as the base value of the construction design. This additional data from the census allowed the model to fit better to the tested data.



Prediction Error Evaluation

The model provided valuable feedback on its ability to predict. Two main metrics were focused on which are the Mean Absolute Error and the Median Absolute Error. They both share the same units of the housing price and are simple to understand. Independently they introduce issues related to the presence or absence of outliers. The mean absolute error has issues related to outliers where the median absolute error can ignore them both to the advantage and detriment. The median absolute error gives a clearer estimate of how far off the majority of the predictions are with the majority of the housing. The difference between the mean and the median indicates how much outliers impact the majority and how far off the median is in predicting more expensive housing options. The mean absolute error is more tolerable of deviation from the housing for more expensive housing than for less expensive housing. Since these two values differ it indicates that the model is very good at most housing price predictions but can still handle outliers within a reasonable range of error. The Median absolute error for the model was \$29,903 and the mean absolute error was \$53,318.



III. Python Notebooks

Colab link to notebook: [link](#)

IV. Discussion Responses

Below are our final answers to the case discussion questions:

After reviewing the data and the type of problem a supervised learning model would be the best option. This is because of the clear prediction goals and the fact that the data can be labeled clearly such as square footage, location, bedrooms, and etc.

For handling the insurance requests, the zip codes for the most part already reflect lower or higher housing prices naturally. However, this does raise questionable ethical reasoning but a big concern is the current rise of legal investigations going into price fixation. Being involved in an investigation from the FTC about fixing housing prices in these areas may result in more than just a federal fine but a possible housing price adjustment reaction.

V. Actionable Items

After doing further research looking at census data and analyzing datasets for other areas some value information to start collecting in the data would be specific features including kitchen quality, living room grade, and master bedroom grade. Further information to also include is how fast housing options sell.