## **Predicting the sale price of a home with basic realty data**

DSC 550

Allison Forte

March 4, 2023

## **Introduction**

Homes are bought and sold every day. Homeowners rely on knowing the value of their home for decisions relating to their home and family’s financial future each day. Websites exist that calculate home value but what happens if I change a feature of my home? How much will my home value increase if I add 650 square feet? What if that square footage includes a bathroom? Another bedroom?

A tool that the average homeowner can access and manipulate puts more fact-based decision-making in the homeowner’s hands. Being able to rely on data to tell you how your home’s value would change given work you plan to complete or weighing a home addition against the prospect of moving would help the average homeowner have confidence in their decisions. While looking at comparable houses can have value, adjusting the facts for your own home would be more valuable and would be a better guiding figure.

This initial project has been completed using public data obtained from Kaggle. The dataset obtained is specific to King County, Washington. This work could be expanded to homes outside of King County, Washington by obtaining a larger data set that is not specific to one county.

The dataset contains over 21,000 entries and included data about sale price, size, number of bedrooms and bathrooms, square footage, year built, quality of the home, the view from the home, and even year renovated, if applicable. This dataset will allow for training and testing data though the results are limited to King County, Washington. Additional data from other counties could be put through this same process.

## **Business Opportunity**

Opportunity exists through this project for the homeowner as well as other potential groups. Home improvement contractors, in particular, could use a tool resulting from this project to make a case around home renovations. If they know additions will add value to the future sale price, they could possibly convince more homeowners in their area to take on larger renovation projects. Having data from every area of the country would be important for a fully useful tool.

## **Analytical framework**

First, this project started with the creation of an idea and finding data. The decision was made to start with data specific to one county to see if the idea was worth further analysis. The exploratory data analysis showed that there were some outliers that would need to be handled before creating a model. Homes with extremely large sizes and extremely high sale prices will throw off the average pricing and make model creation unsuccessful. Being able to create a model using only single-family residences could help avoid some of these outliers that likely come from commercial properties.

Next, steps were taken to prepare the data. Unnecessary columns were removed, as were outliers. 372 rows were removed after being labeled as outliers. The data was checked for missing or invalid entries. Duplicates were also searched for. None were identified. Knowing a regression model would be used, several features were turned into categories and then turned into dummy variables. Specifically, Grade of the home and View were turned into dummy variables.

The final part of preparing the data for model creation was checking that all prices were numeric. Once this was confirmed, the data was split for training and testing, reserving 20% of the data for testing.

With the data prepared, a linear regression model was fitted with the data. This initial model accounted for a fair amount of the variance in the data but included every feature that was included in the dataset. Rather than rely on this model, a PCA model was created. The features were filtered to keep the features that accounted for 90% of the variance. By limiting the features included in our model, we were still able to account for a fair amount of the variance in the final data but now only using 12 features instead of 20. We relied on R-squared to compare models to each other we well as root mean squared error (RMSE) to see how much error we had in the models.

A final model was tried without success. A variance threshold model was attempted, but limiting the features to the ones with the most variance produced a model that accounted for significantly less of the variance in the data. Given this model was limited to only 5 features, it is not surprising that it did not account for a significant amount of the variation in the data.

## **Conclusion**

There is potential in this project with additional data. A model was created that accounts for a fair amount of the variance in the data using the 12 most influential features of the data. However, the model building completed in this project did not account for as much of the variance in the data as expected.

It seems likely that a better model could be built using a bit more time and information. The data set used included information about the latitude and longitude of each home. This data was transformed into a geohash but unfortunately, could not be used in the model in the way it was created. Additional time spent incorporating the geohash data into a model could lead to a more useful model.

Adding in realty data from other county’s across the country could also make a more robust model. The current PCA model was created only using one location’s data and therefore can only be used for other homes in that area. To create a truly useful tool, data would need to be added from other counties. Perhaps this would lead to using the county as one of the features in the model rather than the specific location of each home.

While this project could be useful for homeowners and contractors, it is likely that real estate agents would see little value in it. They have tools they already use to determine sale price of a home and would likely not see tremendous value in adding a different tool. One area that this does have potential to help the real estate community is in valuing homes that have had recent work done. Perhaps if a model could be created that accurately predicts sale price of homes that specifically have had a renovation, then they would see value in it. Getting buy in from the real estate community could lead to significantly more resources for the project. Data sets from across the country could be pooled and used to train a model that works across the country.

With the right data and a more robust model, this endeavor does have potential success ahead. Tailoring the end product to those who stand to benefit from it the most will be important. Giving homeowners another tool to help them make solid financial decisions has the potential to help countless families if it is put to use in the right way and is reliable.