Colin Petersen STAT 1361 Final Nontechnical Report

For this assignment, I was given a dataset with information of over 1,000 homes in Pittsburgh and Virginia. I used this dataset to create and test multiple models that aim to accurately predict the price of any home. I compared these models to each other and determined the best model. In my analysis, I discovered a few indicators of home price that can be exploited to identify undervalued or overvalued homes, and I discovered the context in which my numerical model is most useful.

First, houses in Richmond are, on average, more expensive than houses in Pittsburgh. Almost all the models that I built identified that a home in Virginia will sell for more than a home in Pennsylvania. This is confirmed by the fact that in the data I was given, the median home price in Richmond is over $100,000 more than the median home price in Pittsburgh.

In addition to state being important, almost all of my models indicated that the square footage of a home is significant in predicting its price. In fact, square footage was consistently identified as the most important predictor in the final “best” model I used to predict price. As square footage increases, price tends to increase as well. Perhaps this can be used to identify undervalued homes in Pittsburgh and Richmond. If a very large home is worn down but incredibly cheap, it is possible that renovating the home could provide a substantial profit. This does need further investigation, however.

Interestingly, it seems that the age of a home was not identified as being very important across all quality models. This statistical trend goes against intuition, as most would assume that home buyers want to purchase homes within a certain age range. Maybe consumers do not care about the age of their home if it is in good condition. Maybe some consumers prefer historic homes and some consumers prefer more modern homes, and these consumer preferences even out. Speculation aside, it might be valuable to do a consumer preference survey in the future to determine how home buyers make decisions about the preferred age of their future home.

Another key point from my analysis is that predicting home price accurately can be difficult due to the wide range of prices that any given home could be. Home prices in the dataset I was given range from around $35,000 to almost $4,000,000. The model I created tends to lowball the value of really expensive homes, while being relatively accurate in predicting the value of normal homes.

It is safe to say that the final model I built is not perfect in predicting price. My model is best used as a tool to help a knowledgeable realtor to predict price, not as a standalone method. Let us say that the theoretical “true” price of a home is $400,000. If my model identifies that home to be worth $500,000, PA-VA realty might try to purchase that home at $450,000 and lose money. The model I created is still a valuable tool to get a frame of reference for the significant predictors of price, and to make a rough prediction of a home’s true price before it goes on the market. That said, the model can only make predictions on measurable and recordable things like square footage and the type of roof a home has. The amount of natural light, the neighborhood vibes, and other features and trends are difficult to capture statistically. Additionally, my model tends to be much more accurate when it predicts the price of a home to be $500,000 or less, and much less accurate when it predicts a home price over that amount. For these reasons, my model should best be used as a tool in combination with other qualitative analysis. It is not a complete and decisive system for determining the price of a home.