This post shares a modeling case study. It is an example of how exploratory modeling with very limited data can can yield very accurate guidance for decision-making. I chose this real estate problem out of personal interest but also because we are branching into working with public domain real estate and geolocation data. The problem is to predict home sale price in an interesting U.S. zip code using just basic data from a realtor’s quarterly newsletter -- number of bedrooms and bathrooms, living space square feet and categorical property size (e.g. huge, medium, small).

The 45243 zip code in suburban Cincinnati OH USA covers an almost crazy range of affluence. Here are pictures of houses changing hands over just the past 90 days. How could a simple model predict real estate prices within +/-10% to 20% accuracy across such a crazy span of demographics? Mathematically, data that span multiple decades are a good challenge. A second, possibly overlooked challenge is, once we have an exploratory model, how can we render it in simple form such as a spreadsheet calculator for a non-data expert to digest? How do we structure and curate that such that our model is easy to extend it with future data? We could call this simulation model the “manager test,” but for the sake of this case study, we will call it the “non-coder spouse test.” Practically, I invested about a half day of effort starting from a local realtor’s quarterly newsletter and its clickable links and pictures.

A picture containing grass, tree, outdoor, building

Description automatically generated A house with a flag on the roof

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generatedThe model is based on the U.S. 45243 zip code that has a common denominator of excellent public schools and thriving communities that are a sought-after place to live. One, largely residential village, Madeira OH, contains essentially all the zip code’s retail business. Indian Hill OH and the Sycamore Township area served by Indian Hill public schools is home to corporate executives and scions of Cincinnati’s pro sports teams. Local shops served the late Neil Armstrong who resided in 45243 and taught Aerospace Engineering at the University of Cincinnati. A Madeira trend is rebuilding original 1950’s and 60’s houses into more expensive homes.

The regression modeling example is based on a dataset of 74 home sales from a realtor’s newsletter circa January 2022. The sale prices (in thousands) range from $160K to $2.4 MM –a 15X range. However, log(price) creates a less-skewed distribution for modeling.

The plots or actual versus model-predicted below compare regression results for models where Y is Price\_Sold and log(Price\_Sold). In both models, Log of Square feet was used as an X because of its large variation in the data set. Lot Size was modeled as a three level, nominal variable: Less than 1 acre, 1 to 3 acres and greater than 3 acres. Renovated\_Ext is also a nominal “yes/no” variable representing a judgement call on whether the house has a renovated exterior or not. This makes sense because it is an obvious differentiator within the data set.

The model for Log(Price\_Sold) is superior across the range of prices.

|  |  |
| --- | --- |
| **Actual by Predicted: Y =Price\_Sold** | **Actual by Predicted: Y =log(Price\_Sold)** |
|  |  |
| **Regression Effects**  Baths (p=0.0027) [Continuous variable]  Log(Sq\_Ft) (p<0.0001) [Continuous variable]  Lot\_Size (p<0.0001) [Nominal variable]  Renovated\_Ext (p=0.26) [Nominal variable] | **Regression Effects**  Baths (p=0.0072) [Continuous variable]  Log(Sq\_Ft) (p<0.0001) [Continuous variable]  Lot\_Size (p<0.0001) [Nominal variable]  Renovated\_Ext (p=0.027) [Nominal variable] |

To use the log(Price\_Sold) model, it is necessary to back-convert the predicted Y to price as exp(log(Price\_Sold). 80% of the predicted prices are within 20% of actual prices. This is a strong result given the span of prices covered by the model. There is one outlier where the predicted prices were $626K and $474K in the two models, but the actual price was only $290K. There is obviously a special cause in this situation –an effect that is not included in the model such as a distressed sale or a serious issue with the house in question.

**%Error in predicted price from log(Price\_Sold) model**

**A graph with numbers and a line

Description automatically generated with medium confidence**