The Effect of School Ratings on Real Estate Property Values

It's deceptive and if you are buying a house in Indian Hills subdivision in Cobb County, Georgia, not knowing the market could cost you. Indian Hills is a large suburban subdivision built around a golf course and an amenity package. The subdivision is unique in that straddles two school districts. One of these is East Side Elementary, Dickerson Middle School and Walton High School. The other part of the subdivision is districted to East Valley Elementary, East Cobb Middle and Wheeler High School. These are two very different districts and not thought of as being equal in education quality. This project will investigate how the assigned school district affects the final sales price of the home.

Indian Hills was one of the first planned unit developments in Cobb County. It was begun in 1969 and originally planned for 500 homes. Today it has grown to 1680 homes on over 2,000 acres. Due to its very large size, there is always an active real estate market generating data. Currently there are over 40 listings and within the last year, I was able to identify over 30 home sales which were evenly split between the two school districts. Home prices in the subdivision tend to be between 300,000 and 500,000 making it an ideal target for young families with children. So, school district should be an important factor in home buying decisions for most buyers.

I could have used sales data from a much wider area, but I thought it was important that the homes be relatively close together so that locational differences are minimized and that they be roughly similar in age, appearance and size. I was willing to sacrifice quantity of data over quality of data.

This information will be important to home buyers, especially ones that are not familiar with the area. It will also be useful to real estate appraisers when doing work in Indian Hills.

Data Identified and Proposed Methodology

I will be using school ratings from Great School Ratings, US News and World Report, Niche.com and School Digger to quantify the difference between the districts. I will give equal weight to Elementary, Middle School and High School ratings. This data is categorical (A, B C for example), based on ratings (#5 out of 40 for instance) and numerical 9 on a 10 scale. To generate an average rating, the data will need to be normalized.

Real Estate closing data will come from Zillow.com which in turn gets its information from the two large multiple listing services in the Atlanta area, namely MLS (Multiple Listing Service) and FMLS (First Multiple Listing Service). I won't be able to inspect the homes so remodeling, curb appeal and view will generate some noise in the data. Since the subdivision is based around a golf course, there may be some houses with a golf view. I suspect these will appear as outliers and be easily identified. I found some ward and school boundaries geo data on the Cobb County web site but I don't know if these will be needed since the real estate data already identifies the school districts.

Of course, this problem lends itself to regression if the school ratings can be quantified. It will be interesting to see if the best fit is linear or non-linear.

Results

I was able to import the data into an IBM Watson project. I discovered that Watson has some very good tools for data wrangling with a CSV file. Thus, you won't see any code in the Jupyter Notebook cleaning up the data file. As suspected, the problem lent itself to regression analysis namely multiple regression analysis as we discovered in one of the earlier courses examining home sales. The multiple regression results were both surprising and disappointing. Disappointing because some of the coefficients were negative, namely for bedrooms and unfinished basement. This is counter intuitive since I would think most people would prefer more bedrooms and unfinished basements. I was not surprised to see square footage of the home has a large effect on price and the highest correlation at .57. I was also disappointed that the data did not yield a high actual vs fitted value for price with only 54% of the price explained by the multiple linear regression “multi-fit”. What was surprising was the limited impact the school district had on price. Although positive, I expected it to be greater. The square footage coefficient was 5.94 vs 1.44 for school district.

Discussion

Although I was willing to sacrifice data quality over data quantity, I think the project suffered from a small data sample. Although I restricted the sample to sales over the past 12 months, to avoid having to make price adjustments, I wonder if there would be a way to include data over a 36-month period by making pro-rated price adjustments based on overall price increases.

I also think that with more data, subjective things like curb appeal, seller motivation, view and perhaps even remodeling would have created less noise in the data. I think this is the reason some of the coefficients were negative and the expected vs actual values were not closer.

Conclusion

I think the school rating data clearly show that the two school districts studied are very different in perceived quality. While I was pleased with the similarity of the homes in the data set, I expected there to be a higher effect of school districts on price. The end result of this project is to disprove something I believed to be true with quantify-able data science. I think herein lies the true value of data science.