Predicting Property Value Using An Artificial Neural Network

William Hammond, Harry Lowell

Research:

Problem: We are attempting to build an Artificial Neural Network that can accurately predict changes in housing value in the City of Rochester.

Previous Work: Similar works has been done by a number of universities and corporations. These universities include, Case Western, Middle Tennessee State, Georgia Institute of Technology, Lincoln and many more. The Commercial Broker Association also has conducted their own research.

Challenges: Housing models are better when they are restricted to sub markets, using national housing data doesn't build an effective model [1]. This is due to a number of region specific factors, mostly socio economic factors. Since our model is going to be restricted to the Rochester area, there will be a number of instances where high crime areas border areas mainly inhabited by college students, so going one block one way or the other could drastically affect housing value. Different properties are also valued in different ways, so there needs to be a method in place to restrict our data to exclusively residential properties. On top of this, the housing market is also affected by season, so we also need to account for the time of the year in our model [1]. All things considered, the greatest challenge of this project will be to create a model that

accurately accounts for the qualitative data that must be taken into consideration to accurately determine market value.

Ethical Considerations: The biggest ethical consideration of building a model like this is the possibility of being able to predict areas of gentrification. Although it is unlikely that this model in particular would be able to do so, research is being done to build more advanced models. Gentrification is the process of people of a higher socio-economic class moving into poorer areas due to lower housing cost. As more and more wealthy people move into these areas housing prices rise, forcing the poorer people out of their homes. If a model could predict the area of gentrification, speculative investors could buy up chunks of housing in these areas and target their marketing such that gentrification happens more effectively and quickly. Two things have to be taken into consideration, the first of which is whether or not we consider gentrification in and of itself to be an issue and the second consideration we would have to make is whether or not the targeted pursuit of it to be an issue. All prediction models inherently have error, so a model of this nature may cause an area to be gentrified that otherwise would have been left alone.

Business Case: The business case for this project is straightforward. Every developed nation has a vested interest in real estate so this is a problem that is trying to be solved everywhere. This model could be used by individuals searching for new homes or for broker agencies to identify areas of growth or decline. Models such as this are useful for reducing the amount of subjective determination that arises from using traditional appraisers [2]. If the model would be able to

predict the growth of real estate it would allow buyers to know where it would be best to make their investments in property to get the most return on their investments by buying those properties it predicted will have the most growth.

Data Issues: There is a varied list of issues that could be experienced while working with real estate data. One issue that we are certain to experience will be the issue of the layout of certain property listing. Since Rochester is not as well planned out as a city like New York City, we will have to be aware of the varying layouts. Another common issue in regards to real estate data is the decision on which feature/s will provide the best results, real estate data has a large number of features so this could be a lengthy process to determine which is/are the best to work with. At first glance, finding enough listing data to sufficiently train a neural network seems like it would be a challenge but a number of different studies have shown that limited training sets could be used effectively [8].

Other Data Considerations: The special case and outlier data for property value are more well defined than in other domains. We can easily identify outliers by their status of being in "Foreclosure", "Short Sale", "Corporate Owner" or "Lender Owned" [1]. To build a model better suited to the average buyer, we should exclude well defined outliers such as these. Another consideration we will have to make in regards to our data will have to be the various types of features that our data will contain(binary, continuous, etc.), as this could potentially cause errors in our calculations, thus it is something that we will have to be checking frequently. In regards to the data we will be using, we will fortunately be working with thoroughly cleaned data due to the nature of real estate data.

Previously Used Algorithms: Traditionally, hedonic regression has typically been used to predict real estate value [6]. A hedonic model breaks the property down into features such as number of bedrooms, square footage and the value of the constitution pieces are valued based on the demand in the market. The hedonic models don't often capture all of the qualitative features that affect the value of a property [2]. To account for this, researchers have been turning to the use of Artificial Neural Networks (ANNs). In either case, feature selection is more important than the model itself. There are a wealth features to choose from, so the most useful features need to be chosen. Stepwise and Best Subset Regression is often applied in order to determine where features are best suited for prediction. Once these features are identified, a model is built around these features to predict the response variable, typically market value. Most models employ the classic using of hedonic features but new work is being done with non traditional features, namely search engine data [3]. Search engine data has been shown to predict as good or better as hedonic models [4]. It also has the advantage of being constantly aggregated, so the model could more accurately represent rapid changes in the market. The Hedonic model also tends to not reflect the geography of the unit [4].

Other Issues: A lot of useful features for this project, namely features based on proximity such as proximity to city center, downtown, shopping districts, major highways etc will need to be calculated. Different features are more meaningful in different areas so we will need to spend some time doing regression to identify what features are the most meaningful in Rochester. Hundreds of different features are used in predicting market value so we are going to need to try

multiple different sets to determine what set is the most meaningful. There are also a number of outside factors that contribute to housing prices. These features are what were responsible for the crash in 2008 so even if our model can accurately predict housing value based on the current state of affairs of the world, we don't know what changes could drastically influence housing prices, or if it's possible to create a model to in any way account for the possibility of sudden changes in housing values. There is some current research that search data may be able to predict market crashes 6 months to a year before the crash occurs so we may find a possible solution to this issue there [3]. We also need to decide exactly what geographic region of the greater Rochester area we choose to use. Do we restrict our data to the city of Rochester or do we include Chili, Penfield, Victor Webster etc? According to New York census data, Rochester has a median income of $32,382 which is $20,00 below the New York average and 32.9% of its inhabitants live below the poverty line [7] whereas if you look at Monroe county, the median income is $52,394 with 15% of people living below the property line. So there is clearly pretty wide income differences even in this limited geographic area, which will be reflected in the housing prices.

Distance Metrics: We will begin by looking at the L2 norm. Although the city of Rochester is planned, properties of the L2 norm are still useful. The L1 norm would be more accurate when dealing with proximity based features, but computing the L1 norm would be tedious and because of the fact GPS coordinates are readily available computing the L2 is easy and fairly accurate [5]. Some of the features chosen will be binary so there may be instances where the hamming distance is useful.

Method: Our hypothesis is that crime rates and proximity metrics will account for the greatest difference in property values in Rochester. We will build a set of features accounting for the bread and butter hedonic features, such as square footage, number of rooms etc supplemented with different possible proximity metrics and social metrics. We know that these three sets, hedonic, proximity and social, are all useful in varying degrees based on what geographic region you are looking at. Due to the relative ease of implementation and due to the fact that it has shown better results in recent years, we will use an Artificial Neural Network as opposed to a regression based model. There are some considerations that need to be made when working with ANNs. First of which is its black-block nature. Because of this we are unable to inspect its structure to determine potential flaws. To account for this, some guess and check work needs to be done to gauge the model effectiveness. This is easy with modern computing libraries. As far as feature selection goes, there are features that are universally good at predicting market value such as square footage, number of bedrooms, what neighborhood etc. These features will be used as a base and we will select additional features initially based off our intuition and knowledge of Rochester.

Validation: We will train and validate the model using historic data so we can compare our predicted values with actual listed values. Because we are using ANNs, some guess and check work needs to be done to choose the model. There are different initial settings and minor calibrations we can do to the model. To determine which of these configurations is the most effective, we will need a metric to measure the quality of the model. To do this we will use the

SSE where we will measure the euclidean distance between our predicted value and the actual value. The residuals of the features can be used to gauge the effectiveness of individual features.

Algorithms To Use: Artificial Neural Networks, Squared Sum Error, L2 Norm

Report:

Algorithms Used: Artificial Neural Network, PCA, min-max normalization, z-score normalization

Challenges: This project had more challenges than were predicted when the original idea was formed. Originally we thought that our data would be fairly easy to obtain since there were numerous resources that we would be able to reach out to and ask for access to the data we were looking to perform our prediction on. However, this idea of being able to easily access the data we needed was quickly removed from our minds. In the end we had to change the overall objective of the project to fit the data that we had access to and would be able to run our algorithms on. Instead of relying on primary attributes we had to turn to using statistics collected by Zillow and offered as mass data dumps. We were forced to pivot into using features that told us more about the health of the market, like the % of homes sold, % of homes that had to have their listing price reduced. After to a little bit of research about the topic of acquiring data it would seem that it is not uncommon for real estate agencies to not be very willing to share the data that they have with others. It is entirely possible that because we would just be using the data for a class project that the companies we contacted realized that there would be no chance for a return of investment and thus decided that we were not worth the time to provide us with the data that we needed.

What was interesting: When we started reading in the data regarding the housing market, we would plot out each of the features that we were using. What was interesting about this was that,

in the plots of certain features you could clearly see drops in the data where the housing crash of 2009 occurred. All of the different features reflected the crash, and the crash affected all the county, state and national data we looked at. The fact that we were training off of a time series was generally interesting. When we initially ran our model and plotted the error it looked like that the error for different data sets all were time dependent. No matter what county or state we looked at the patter was the same. We later found that this was caused because the predicted value was the same for all input so it was just reflected the actual median value of the over that set of time. Although this won't be particularly important to our results, it opens up some interesting questions. Instead of treating our data as a time series we could shuffle the data to try and make the model more generic. There also may be other regression models that better handle time dependent data.

Another thing that was interesting about this was that right before the immediate crash, there was a sizable increasing in the values, making for a very drastic point of inflection. The way neural networks behave when improperly was also strange to see. Our model was about as wrong as possible and we had better results than when we had a proper model. Although it was blatantly obvious during the validation step that we had made some simple mistakes, I could see how if this stage of development wasn't taken seriously there could be serious repercussions.

The power of R's neural network library was also incredible interesting. By just formatting a data set properly we were able to pass it into a function call and get a trained model. The function gave us the use of 20+ parameters that could be tweaked inorder to better fine tune the model.

What did we learn: The earliest lesson that we learned was that it is not as easy as you think it will be to acquire the data that you need. Despite the fact that we reached out to a number of organizations including the National MLS, Local MLS, County Clerk, Zillow and the Democrat and Chronicle we were only able to talk to a single representative. Due to the politics of the real estate industry, listing data is very hard to obtain. We were operating under the assumption that because there was so much published research on the subject, finding data would be easy. In hindsight we definitely should have done research and figured out if we were going to be able to get access to the data that we needed to use for our project.

Overview of the data: For this project there was little need to perform any cleaning on the data that we used, this is because of the nature of the real estate that we were using it would have to be cleaned prior to use by real estate companies. This ultimately saved us time in the process of using our data. The data is offered through Zillow, a real estate database company. Although they keep their primary data (historical listings, square footage, etc…) they offer a large amount of secondary data (median home price per neighborhood) that we used to train our model. In regards to the data that was available via Zillow, we used multiple data sets of varying types of data, including percentage of homes that were sold for gain, percentage of homes in a given region that increased in value, median of the price reduction for homes in a region, to name a few. For each of these data sets, the features that we used were the the months of October 2010 to September of 2015. We were restricted to this date range because some of the features we restricted to this range.

Bibliography:

1 - Corisini, Kenneth Richard. *STATISTICAL ANALYSIS OF RESIDENTIAL HOUSING PRICES IN AN UP AND DOWN REAL ESTATE MARKET: A GENERAL FRAMEWORK AND STUDY OF COBB COUNTY, GA*. Thesis. Georgia Institute of Technology, 2009. Atlanta: Georgia Institute of Technology, 2009. Print.

2- Hamzaoui, Y., & Perez, J. (2011). Application of artificial neural networks to predict the selling price in the real estate valuation process. In *2011 10th Mexican International Conference on Artificial Intelligence* (pp. 175 - 181). Piscataway: IEEE.

3- Wu, Lynn and Brynjolfsson, Erik, The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales (August 30, 2013).

4-Limsombunchai, Visit. "House price prediction: Hedonic price model vs. artificial neural network." *New Zealand Agricultural and Resource Economics Society Conference*. 2004.

5-Dubin, Robin. "Predicting House Prices Using Multiple Listings Data." *Predicting House Prices Using Multiple Listings Data*. Volume 17, Issue 1 ed. Vol. The Journal of Real Estate Finance and Economics. Kluwer Academic, 1998. 35-59. Print.

6- Peterson, Steven P. and Flanagan, Albert B., Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. Journal of Real Estate Research (JRER), Forthcoming.

7- U.S. Census Bureau: State and County QuickFacts. Data derived from Population Estimates, American Community Survey, Census of Population and Housing, County Business Patterns, Economic Census, Survey of Business Owners, Building Permits, Census of Governments Last Revised: Wednesday, 14-Oct-2015 16:29:10 EDT

8- MORANO, PIERLUIGI, FRANCESCO TAJANI, and CARMELO TORRE. *Artificial Intelligence in Property Valuations An Application of Artificial Neur Al Networks to Housing Appraisal*. Bari: Polytechnic of Bari, 2014. Print.