Data Mining: An Empirical Application in Real Estate Valuation

Ruben D. Jaen

Florida International University
University Park, PC236
Miami, Fl 33199
jaenr@fiu.edu

Abstract

This paper presents the insights gained from applying data mining techniques, in particular neural networks for the purposes of developing an intelligent model used to predict real estate property values based on variety of factors. A dataset of over one thousand transactions in real estate properties was used. The dataset included 15 variables obtained from the multiple listing system (MLS) database and captured information on transactions taking place during a period of three years. The results from applying data mining techniques to predict real estate values are promising. Future plans and recommendations for further expanding the study are given.

Keywords: Data mining, real estate valuations, home appraisals

Introduction

The factors that determine housing prices are of interest to urban planners, developers, real estate professionals, and financial executives as well as most American homeowners. According to a 1998 Federal Reserve survey (Kennickell, *et al*., 2000), 66.2 percent of U.S. households are homeowners and housing investment amounts to 33 percent of household net worth. The number of new home sales as well as home resales are an important component of the U.S. economy and data concerning these transactions is closely tracked for the purpose of gauging economic activity and formulating appropriate monetary and fiscal policies. This paper examines the factors that determine housing prices in a sample of over 1000 home sales in Miami-Dade County during the period of 1999-2001.

Sales of homes take place in the marketplace dictated by the usual rules of supply and demand. Since this is not a perfect market, there is a great latitude for judgment in arriving at the selling price, thus the job of a real estate appraiser has been described as more art than science (Gilbertson, 2001). There are three well known approaches for estimating real property value (Gains, 2001): a) comparable sales method, b) cost method, and c) income method. In theory, an appraisal report uses all three approaches to estimate the value of a property. Based on the type of property, a degree of priority is assigned to each method used.

Basically, none of the three methods is foolproof, and in the end the appraiser has to make a judgment call. This study uses knowledge discovery techniques such as neural nets and decision trees to examine the dataset on real estate transactions to identify factors that influence selling price and proposes the development of a model that can be used to predict real estate prices.

Neural Nets

Neural nets are a class of predictive modeling system that work by iterative parameter adjustment. Structurally, a neural network consists of a number of interconnected elements (called neurons) organized in layers which learn by modifying the connection strengths (i.e., the parameters) connecting the layers. Neural nets usually construct complex equational surfaces through repeated iterations, each time adjusting the parameters that define the surface. After many iterations, a surface may be "internally" defined that approximates many of the points within the dataset.

The basic function of each neuron is to: (a) evaluate input values, (b) calculate a total for the combined input values, (c) compare the total with a threshold value and (d) determine what its own output will be. While the operation of each neuron is fairly simple, complex behavior can be created by connecting a number of neurons together. Typically, the input neurons are connected to a middle layer (or several intermediate layers) which is then connected to an outer layer. To build a neural model, we first train the net on a "training dataset", then use the trained net to make predictions. We may, at times, also use a "monitoring data set" during the training phase to check on the progress of the training.

Each neuron usually has a set of weights that determine how it evaluates the combined strength of the input signals. Inputs coming into a neuron can be either positive (excitatory) or negative (inhibitory). Learning takes place by changing the weights used by the neuron in accordance with classification errors that were made by the net as a

whole. The inputs are usually scaled and normalized to produce a smooth behavior.

During the training phase, the net sets the weights that determine the behavior of the intermediate layer. A popular approach is called "backpropagation" in which the weights are adjusted based on how closely the network has made guesses. Incorrect guesses reduce the thresholds for the appropriate connections.

Neural nets can be trained to reasonably approximate the behavior of functions on small and medium sized data sets since they are universal approximators. It is well known that backpropagation networks are similar to regression.

Rule Induction/Decision Trees

Rule induction is the process of looking at a data set and generating patterns. By automatically exploring the data set, the induction system forms hypotheses that lead to patterns. The process is in essence similar to what a human analyst would do in exploratory analysis. For example, given a database of demographic information, the induction system may first look at how ages are distributed, and it may notice an interesting variation for those people whose profession is listed as professional athlete. This hypothesis is then found to be relevant and the system will print a rule such as:

IF Profession = Athlete
THEN Age < 30

This rule may have a "confidence" of 70% attached to it, indicating that 70% of the athletes were in the age group of 30 years or less. Decision trees successively partition a data set based on the relationships between predictor variables and a target (outcome) variable. When successful, the resulting tree or rules indicates which predictor variables are most strongly related to the target variable. It also describes subgroups that have concentrations of cases with desired characteristics. Decision trees such as C&RT are entirely non-parametric and can capture relationships that standard linear models do not easily handle.

Data Set

A data set was obtained from the MLS (multiple listing service) containing information on 1229 transactions involving residential real estate properties sold within the period of 1999-2001 in the city of Coral Gables, Florida. In addition to descriptive information regarding property address, the following numeric and categorical variables were included in the analysis.

Area (location of the property)	Categorical
Lot square feet	Numerical
No. Bedrooms	Numerical
No. Baths	Numerical
Living area sq. feet	Numerical
Style	Categorical
Year built	Numerical
No. Garages	Numerical
Pool	Binary (Y, N)
Waterfront	Binary (Y, N)
Assessed Value	Numerical
Property Tax	Numerical
Zip	Categorical
SALE PRICE (dependent variable)	Numerical

After examining the data for outliers it was decided to retain for further analysis transactions involving properties in the range of 100K to 700K. This decision was based on the fact that any properties below 100K are considered extremely rare and are likely to be data entry errors. Similarly at the other extreme, properties sold for over $700K can be considered unique and the usual valuation models do not normally apply to this segment. In other words, there are other intangible factors that come into play. The reduced data set contained 959 cases ranging from a sale price of $108K to $700K with a mean value of $369K.

Data Analysis

In order to have a basis for comparing the data mining techniques with traditional statistical techniques, a stepwise linear regression was conducted to determine the accuracy of the independent variables predicting sales price. Prior to running the analysis categorical variables were converted into dummy numerical variables. Regression results indicate that the overall model significantly predicts sale price, R^2=.78, R^2_{adj}=.78, F(9,895)=345.4, p<.001. The resulting model included 9 variables. This model accounts for 78% of the variance in sale price. The following 9 variables were selected by the regression model as being significant predictors: assessed, sqftliving, pool, baths, year, garage, waterfront and lotsqft.

An examination of the tolerance statistic generated by the model did not indicate a serious problem with multicollinearity. (tolerance >.1).

The next step was to employ a decision-tree modeling technique. In particular the C&RT algorithm was selected because the dependent variable (sale price) is continuous. The algorithm partitions the data set based on the relationships between predictor variables and the target (sale price). The resulting tree or set of rules indicates with predictor variables are most strongly related to the target. At the root of the tree assessed value appears as the single best predictor, followed by square feet and property taxes.

Finally, the neural network with backpropagation algorithm was employed to study the data. A training sample of 50% was used to develop the model and the remaining data was used to validate the accuracy of the model. The resulting neural network consisted of an input layer of 37 neurons, 6 neurons in the hidden layer and 1

neuron in the output layer. The predicted accuracy computed by the software tool was 98%. Although this number is high, it is important to note that for a continuous target variable, accuracy is calculated as follows: 100*(1-absolute value((target value – network prediction) / target value range))). This calculation would tend to produce high values of accuracy whenever the target value range is large, like in the current data set where the range is 592,000. A ten-fold validation method was used by dividing the data set in 10 testing sets. Then for each one of the ten sets, a network was developed using the remaining 9 sets to train the network and the testing set to measure accuracy. The accuracy measures obtained in the ten runs were in the range of 88 to 98 percent.

In order to compute a more realistic and more generic measure for accuracy that can be used regardless of modeling technique, the mean absolute error (MAE) was calculated by computing the absolute error (predicted – observed) for 100 observations in a holdout sample and obtaining the mean. The smaller the MAE, the better predictive power of the model. Table 1 shows the results for the three methods as well as the variables that were identified as good predictors. The variable sqft living was the only one that was selected by all three models. This is consistent with what we would expect to be a main predictor since the construction cost of a home can be estimated using this parameter. Four other variables (assessed, year, half baths, zip) appeared on more than one model.

In terms of MAE, the decision tree produced the best result, while the regression and the neural network produced larger errors. One of the reasons why the Neural Network did not outperform the regression might be that there is not enough complexity in the data set.

Table 1. Summary of Results from different methods.

	Stepwise Regression	**Decision Tree C&RT**	**Neural Net**
MAE	69,396	42,854	71,594
Predictors	Assessed	Assessed	
	Sqft living	Sqft living	Sqft living
	Pool		
	Baths		
	Year	Year	
	Garage		
	Waterfront		
	Lot sqft		
	Half baths		Half baths
		Tax	
		Zip	Zip
			Garages
			Style
			Area

Conclusion

This empirical study demonstrated the applicability of Data Mining techniques, in particular Neural Networks and Decision Trees to identify factors that explain the valuation of real estate properties and developing models that can be used for prediction purposes. The decision tree algorithm known as C&RT produced the best results since the model was fairly simple to understand and produced the least error (measured in terms of mean absolute error) in the test data. Furthermore the decision tree also utilized the least number of predictors (5) to arrive at the solution.

In order to improve the accuracy of the models in future studies it is recommended that additional variables not typically available through MLS, such as local unemployment data, mortgage rates, data on new construction, etc., should also be included. In addition, the use of data mining techniques should be expanded to study other types of real estate properties such as vacant land, income properties, industrial sites and others. Given the tremendous amount of data that is currently available through the internet and other sources, it makes sense for organizations dealing in the real estate industry to try to gain competitive advantage by using data mining techniques to better understand the factors that affect valuation and monitor changes in purchasing patterns.

References

Abraham, J.M. and P.H. Hendershott. 1996. "Bubbles in Metropolitan Housing Markets." *Journal of Housing Research* 7(2): 191-208.

Bartik, T.J. 1991. *Who Benefits from State and Local Economic Development Policies?* (Kalamazoo, Michigan: W. E. Upjohn Institute).

Becerra-Fernandez, I; Zanakis, S. and Walczak S., "*Knowledge Discovery Techniques for Predicting Country Investment Risk*", Journal of Computers and Industrial Engineering, Special Issue on Data Mining, Forthcoming 2002.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. 1998. Discovering Data Mining: From Concept to Implementation, New Jersey: Prentice Hall.

Gains G; Coleman D; Crawford, L. Florida Real Estate Principles, Practices & Law. Real Estate Education Co.; Chicago 2001.

Gilbertson, Barry, 2001. "Appraisal or Valuation: An Art or a Science?" Real Estate Issues 26(3): 86-90.

Kennickel, A. B., M. Starr_McCluer, and B. J. Surette. 2000. "Recent Changes in U.S. Family Finances: Results

from the 1998 Survey of Consumer Finances." *Federal Reserve Bulletin* 86(1): 1-29.

Linneman, P. A., A. Summers, N. Brooks and H. Buist. 1990. "The State of Local Growth Management." Working Paper No. 81. Philadelphia, PA: Wharton.

Ludvigson, S. and C. Steindel. 1999. "How Important is the Stock Market Effect on Consumption?" *Federal Reserve Bank of New York Economic Policy Review* 5(2): 29-52.

Malpezzi, S. 1996. "Housing Prices, Externalities, and Regulation in U.S. Metropolitan Areas." *Journal of Housing Research* 7(2): 209-242.

Malpezzi, S., G. H. Chun, and R. K. Green. 1998. "New Place-to-Place Housing Price Indexes for U.S. Metropolitan Areas, and Their Determinants." *Real Estate Economics* 26(2): 235-274.

Park, S., Piramuthu, S., and Shaw, M., Dynamic Rule Refinement in Knowledge-based Data Mining Systems. *Decision Support Systems* 31:205-222.

Pindyck, R. S. and D. L. Rubinfeld. 1981. *Econometric Models and Economic Forecasts*. New York, NY: McGraw-Hill.

Poterba, J.M. 1991. "House Price Dynamics: The Role of Tax Policy and Demography." *Brookings Papers on Economic Activity* 2: 143-199.

Poterba, J.M. 2000. "Stock Market Wealth and Consumption." *The Journal of Economic Perspectives* 14(2): 99-118.

Rose, L.A. 1989. "Topographical Constraints and Urban Land Supply Indexes." *Journal of Urban Economics*. 26(3): 335-347.

Segal, D. and P. Srinivasan. 1985. "The Impact of Suburban Growth Restrictions on U.S. Housing Price Inflation, 1975-78." *Urban Geography* 6(l): 14-26.

Starr-McCluer, M. 1998. "Stock Market Wealth and Consumer Spending." Federal Reserve Board of Governors, working paper (April).

Von Krogh, G; Roos, J and Kleine, D (etds.). Knowing in Firms: Understanding, Managing and Measuring Knowledge; Sage Publications, London 1998.