Real Estate Pricing Models Using Machine Learning Schemes

Ingrid Rosales, Carlos Hernández
Instituto de Ingeniería Industrial y Sistemas
Universidad Austral de Chile
Valdivia, Chile
ic.rosalesgomez@gmail.com, carlos.hernandez@uach.cl

Abstract

This research presents the analysis and comparison of prediction models for the market price of real states by means of applying machine learning schemes based on artificial neural networks (ANN). Three prediction models are developed, each considering a particular set of attributes and a different ANN architecture. The research is carried out following a 4-stage methodology: analysis, design, construction, and validation. To construct the prediction models, 228 real estates are considered. For each property, 35 attributes were documented. The dataset was split into 2 files. The first one containing 80% of the data for training and testing purposes, while the remaining 20% is left for validation. To reduce the uncertainty, a cross validation strategy was applied. All prediction models are finally compared by means of the error measures MAPE, MAE, and RMSE. In all cases, prediction results present a MAPE between 14% and 16%. In conclusion, the research revealed promising results when machine learning schemes are used to predict real estates' market prices. pricing model.

Keywords

Machine learning scheme, Artificial neural network, Real estate market, Cross validation, Regression analysis.

1. Introduction

The dynamics of real estate markets have always had a profound impact on the sales prices of listed properties, which are influenced by both objective and subjective aspects. Commonly, a property's sale price differs greatly from its tax assessment. Among the objective aspects are the construction materials, area, number of bedrooms and bathrooms, location, etc. Some of the subjective aspects are, on the other hand, architectural style, surroundings, exclusivity, etc. Before buyers get involved in a negotiation it is a common practice to ask advice from a realtor with enough experience to determine a fair price for a property. However, real estate market is often highly speculative and susceptible to market conditions and even to trends. Therefore, sales prices can fluctuate significantly depending on the knowledge of the realtor, the ability to negotiate and the conditions of the local market. The situation of global economy might also have an impact on this sensitive market as the conditions to obtain home loans varies too. For these reasons, it is interesting to explore the potential of artificial intelligence tools that by means of applying models based on machine learning algorithms allows to determine the right sale price of a property.

1.1 Objective

Apply artificial intelligence tools to help determine the right sale price of properties in Valdivia (Chile) by means of implementing prediction models based on machine learning schemes such as artificial neural networks.

2. Literature Review

2.1 Machine learning schemes

Machine learning schemes are algorithms used to find patterns in data sets through experience, without the need of explicitly programming code. There exist supervised, unsupervised, and reinforcement learning algorithms.

In supervised learning algorithms the training is done using labeled datasets that contain the response or class to be predicted. In unsupervised learning, instead, the desired response or class is not known. In the reinforcement learning, on the other hand, predefined actions, parameters, and final values are used.

There are different machine learning algorithms, which are usually grouped as follows:

- Regression algorithms
- Bayesian algorithms
- Cluster algorithms
- Decision tree algorithms
- Deep learning algorithms
- Artificial neural networks

In the study of the real estate markets the hedonic regression models are the most used (Rosen 1974). The theory of implicit prices states the prices of a complex good must be estimated as a function of its characteristics (Morano Tajani 2013).

2.2 Artificial neural networks (ANN)

ANNs consist of nodes or neurons that are combined in an interconnected layered structure. The first level is the input layer, which contains the nodes that receive the externa data. In the second level are the hidden layers that transform the input data for the output layer, whose neurons are responsible for delivering the results generated by the network (Morano Tajani 2013). The ANN topology is determined by the number of layers, the number of nodes in each layer and the transfer function (Hamzaoui Hernández 2011).

The advantage of ANNs is the ability to learn from highly correlated, incomplete, or previously unknown data (Ge et al. 2003).

2.3 Prediction and forecasting

Prediction, or classification, is the determination of the value of an independent variable or class in a set of unknown instances using a model previously trained. Forecasting, instead, is the prediction of the future using time series. Therefore, the difference between these concepts is the time dimension.

2.4 Cross validation

Cross-validation is a technique used to evaluate the results of a statistical analysis and to ensure that they are independent of the split to separate training and testing data. In its simplest version, the dataset is divided into k folds, from where k-1 folds are used for training and the remaining fold for testing. A total of k iterations is performed interchanging the testing fold. Since each iteration will produce different results, they are then weighted (Figure 1).

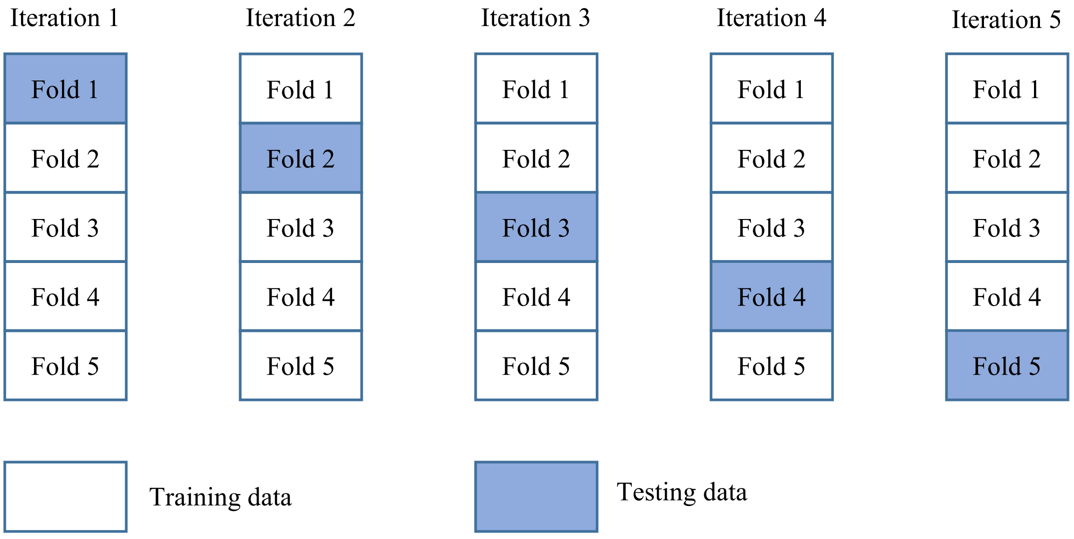

Figure 1. Five-fold cross validation

3. Methods

The research was carried out following a classic 4-stage model: analysis, design, construction, and validation (Figure 2).

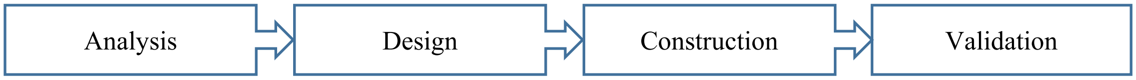

Figure 2. Four-stage model

3.1 Analysis

During the analysis, existing bibliography and scientific articles are reviewed. Abundant information about the real estate market in Valdivia (Chile) is also collected. At this point, the scope of the investigation is defined, and existing software packages are compared to select the suitable one for the research's needs.

No different from any modern city, in Valdivia there are neighborhoods that are more desirable than others. To assess appropriately the value of one square meter of are in each district of Valdivia, the city is divided in 11 zones, each receiving a coefficient to relate its market price to the other zones (Figure 3).

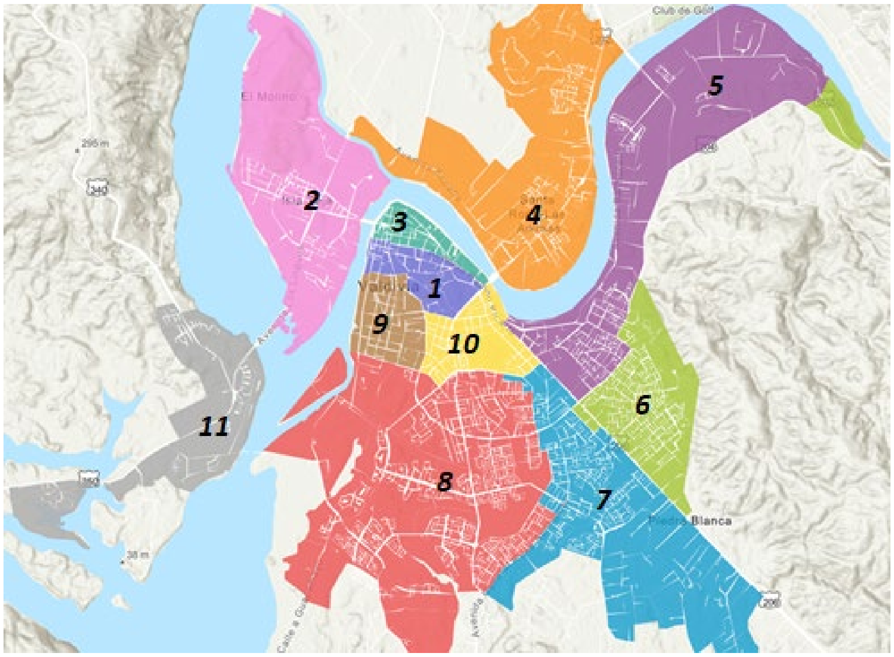

Figure 3. Valdivia's Zoning

The attributes or independent variables that characterize the properties listed for sale are defined during this stage too, being sale price the independent variable or class to be predicted. In total number of 35 attributes are used to create the prediction models.

After identifying all properties currently for sale in Valdivia, a sample of 228 were selected for the study.

3.2 Design

The complete dataset, a matrix of 228 rows (instances) and 35 columns (attributes), is separated into 2 files. The first one, for training and testing, contains 80% of the data. The remaining 20% of data is left in a separate file to be used during the validation stage.

For the purposes of this research, a supervised learning scheme based on artificial neural networks is selected. Using the training and testing file a quick analysis is carried out to determine the correlation between each attribute and the dependent variable, the market price.

Then, from the results of a regression analysis and estimated correlation coefficients, 3 sets of attributes selected from the original 35 attributes are defined to build specific prediction models. The criteria used for attribute selection are based on the level of correlation and the adjusted R^2 value (Table 1).

Table 1. Attributes – Correlation coefficient and model selection

Attribute	Corr. Coef.	Model 1	Model 2	Model 3
Property type	0,380		✓	✓
Area m^2	0,502	✓	✓	✓
Construction m^2	0,813	✓	✓	✓
Bathrooms	0,721	✓	✓	✓
Dorms	0,485			
Floors	0,244			
Front yard	0,125			
Swimming pool	0,301		✓	✓
Architectural style	0,626	✓		
Actual state	0,412		✓	✓
Parking lot	0,057			
Additional construction	-0,296		✓	
Fence	-0,338			
Cellar	0,147			
Barbecue	0,257			
Property for rent	0,069			
Health centers	0,135		✓	✓
Schools	-0,265		✓	✓
Fire stations	-0,167		✓	✓
Police stations	-0,059			
Public transportation	0,115			
Supermarkets	-0,102		✓	✓
Commercial centers	-0,122		✓	✓
Cemetery	-0,044			
Gas stations	0,077			
Industries	-0,176			
Jail houses	-0,156			
Type of district	0,428		✓	
District density	-0,625	✓	✓	✓
Acoustic contamination	-0,092			
Air pollution	0,654	✓		
Exclusivity	0,622	✓	✓	
Location	0,219		✓	✓
Tax assessment	0,847	✓	✓	✓
Tax subsidy	-0,506	✓		

To compare the prediction models, the following error measures have been selected:

- MAPE : mean absolute percentage error
- MAE : mean absolute error
- RMSE : root mean squared error

In the case of MAPE, the interpretation criteria shown in Table 2 is adopted (Moreno et al. 2013).

Table 2. MAPE – Interpretation of typical values

MAPE	Prediction interpretation
<10	Precise
10-20	Good
20-50	Reasonable
>50	Unprecise

In the case of MAE and RMSE, values below a half of the standard deviation are considered poor (Singh et al. 2004).

3.3 Construction

The prediction models' construction carried out completely using the software package Visual Gene Developer and the previously created training and test dataset.

As aforementioned, the selected machine learning scheme is an ANN 2 hidden layers (Figure 4, Figure 5, and Figure 6). To perform the training and testing of the models, a cross-validation technique with 10 folds is used.

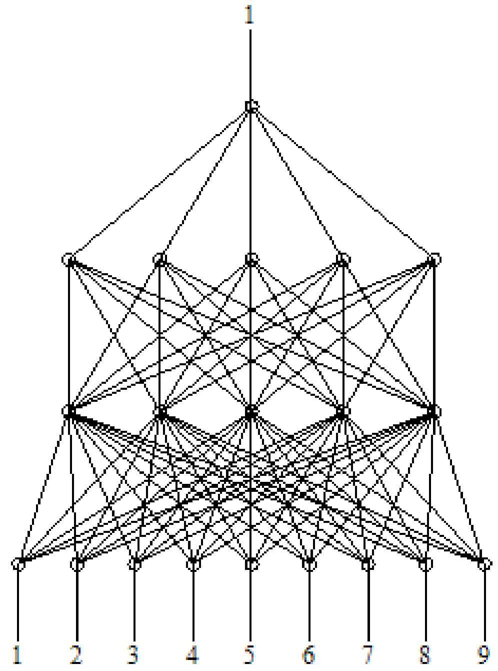

Figure 4. Model 1

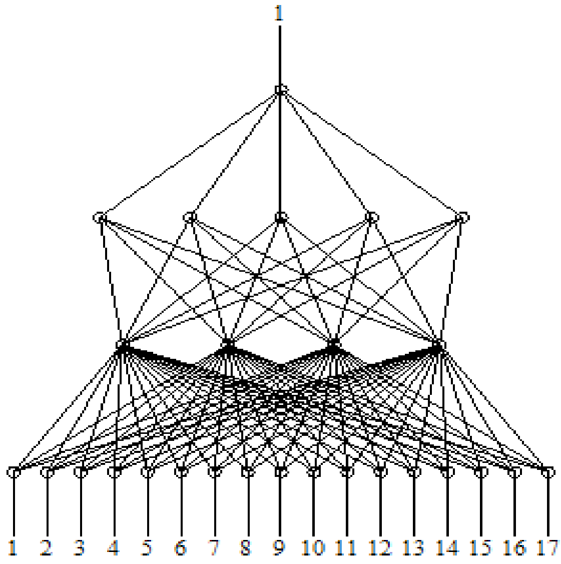

Figure 5. Model 2

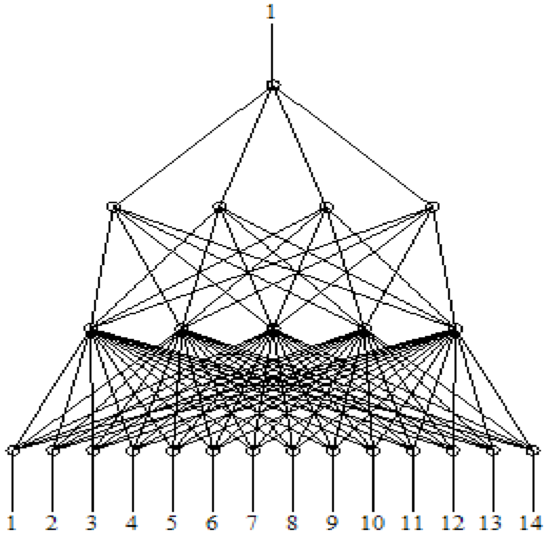

Figure 6. Model 3

3.4 Validation

To perform the validation of the resulting prediction models, the validation dataset (20%). These instances are totally unknown and unseen up to now. Then, from the models' predictions and the actual sale prices the measures MAPE, MAE, RMSE calculated.

4. Data Collection

Actual sales prices and models' prediction for the validation data set are shown in Table 3.

Table 3. Validation dataset - Real estate sale price predictions [CLP $]

#	Actual Sale Price	Model 1	Model 2	Model 3
1	130,000,000	87,129,095	110,679,546	103,631,029
2	233,264,292	175,204,477	150,043,098	124,936,765
3	148,000,000	121,218,074	92,131,391	139,034,258
4	320,000,000	107,087,104	316,402,836	351,801,901
5	189,529,032	169,932,277	129,092,166	126,298,526
6	80,119,091	100,746,823	87,809,684	101,043,834
7	90,000,000	141,466,764	87,148,207	81,757,961
8	25,000,000	64,572,524	57,680,111	51,238,398
9	61,999,828	59,009,768	42,024,614	49,152,037
10	319,998,520	207,769,470	206,425,406	242,438,812
11	85,754,995	71,822,032	69,140,151	67,644,396
12	186,643,309	208,833,595	210,934,405	183,699,781
13	65,032,293	83,180,015	53,972,121	66,738,915
14	68,604,053	72,405,353	74,105,013	68,886,528
15	130,000,000	184,249,373	215,424,666	115,591,902
16	65,000,000	61,274,552	58,845,586	57,430,111
17	215,373,900	209,842,632	193,661,651	236,883,093
18	72,999,691	81,548,434	57,888,473	60,561,943
19	50,000,000	55,630,059	35,206,048	49,633,675
20	70,000,000	74,801,258	50,348,241	53,198,552
21	84,999,751	72,667,511	52,004,014	58,867,154
22	160,000,000	189,206,324	109,885,408	204,455,967

5. Results and Discussion

5.1 Numerical Results

The error measures resulting from the models' predictions with different dataset are presented in Table 4, Table 5, and Table 6.

Table 4. Error measures – Training and testing dataset

Measures	Model 1	Model 2	Model 3
MAPE	10.60	11.42	9.98
MAE	0.029	0.035	0.029
RMSE	0.045	0.048	0.041

Table 5. R^2 – Training & Testing and validation

	R^2	
	Training & Testing	Validation
Model 1	0.950	0.799
Model 2	0.968	0.715
Model 3	0.969	0.711

Table 6. Error measures – Training, testing, and validation dataset

Measures	Model 1	Model 2	Model 3
MAPE	16.247	16.441	14.443
MAE	0.046	0.052	0.044
RMSE	0.074	0.076	0.069
R^2	0.885	0.883	0.901

The resulting error measures are consistent with those obtained by other authors using different methodologies. For instance, a R^2 equal to 0.73 (Saegner 2011), a RMSE equal to 0.6613 and a MAE equal to 0.5135 (Selim 2009) can be found in the literature.

6. Conclusion

Gaining a good understanding of the local real estate market during analysis stage the is crucial for the development of prediction models. To incorporate the preferences of buyers for specific districts, the zoning proved to be a useful tool to compare and weight the sale price of the square meter across the city.

The definition of a wide number of attributes to characterize the listed properties make possible to build prediction models based on different criteria. For instance, both regression and correlation analysis can be used to decide whether to include or exclude certain attributes or independent variables that are not statistically significant.

The availability of separated dataset for training and testing and for validation make possible to determine if the models can predict adequately when a set of unknown data is used. The application of cross-validation technique helps eliminate the bias caused by an unfortunate partition and therefore, it helps reduce the uncertainty of the result by means of performing multiple iterations.

Finally, the resulting error measures from the validation dataset suggest that prediction models based on artificial neural networks can be a useful tool to help estimate the sale price of a property using an objective criterion based on multiple attributes.

References

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy. 82(1): 34-55

Morano, P., & Tajani, F. (2013). Bare ownership evaluation. Hedonic price model vs artificial neuroal network. International Journal of Business Intelligence and Data Mining. 8(4): 340-360.

Hamzaoui, Y., & Hernández, J.A. (2011). Application of artificial neural networks to predict the selling price in the real estate valuation process. 10th Mexican International Conference on Artificial Intelligence.

Ge, J., Runeson, G., & Lim, K. (2003). Forecasting Hong Kong Housing Prices: An Artificial Neural Network Approach. International Conference on Methodologies in Housing Research, Stockholm, Sweden.

Moreno, J., Palmer Pol, A., Sesé, A., & Cajal, B. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. Psicothema.25(4): 500-506.

Singh, H., Knapp, V., & Demissie, M. (2004). Hydrologic modelling of the Iroquois river watersheds using HSPF and SWAT.

Saegner, A. (2011). Determinantes del Precio de Viviendas en la Región Metropolitana de Chile. El Trimestre Económico. 78(4): Páginas 813-839.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic Regression versus artificial neural network. Expert Systems with Applications. 36: 2843-2852.

Biographies

Ingrid Rosales is an Industrial Engineer. She earned a B.S. and Licentiate Degree in Engineering from the Universidad Austral de Chile, Valdivia, Chile. Her research interests include machine learning and neural artificial networks.

Carlos Hernández is an industrial engineer and professor at the Institute of Industrial Engineering at the Universidad Austral de Chile, Valdivia, Chile. He earned a Licentiate Degree in Engineering from the Universidad de La Frontera, Temuco, Chile, a Master of Sciences in Computational Engineering, and a Doctor of Engineering from the Technische Universität Braunschweig, Brunswick, Germany. He is the author of several scientific and engineering articles. He has taught lectures on Discrete Event Simulation, Supply Chain Management, Engineering Economics, Corporate Finances, Financial Engineering, Business Analytics, Data Mining, and Machine Learning for engineering students. He has developed a professional career working for large multinational companies (PricewaterhouseCoopers, BHP Billiton, and Merck Sharp & Dohme). He also worked as a scientific researcher in the Institut für Produktionsmesstechnick at the TU Braunschweig, Germany. His research interests include assembling process techniques, manufacturing process simulation, urban traffic and transportation systems simulation, supply chain design and management, and machine learning for finances. He is a member of IEOM.