Building Models to Predict Real Estate List Prices using Ensemble Machine Learning Algorithms

Carlos Hernández
Departamento de Ingeniería Industrial y Sistemas
Universidad de La Frontera
Temuco, Chile
carlosalberto.hernandez@ufrontera.cl

Ingrid Rosales
Instituto de Ingeniería Industrial y Sistemas
Universidad Austral de Chile
Valdivia, Chile
ic.rosalesgomez@gmail.com

Abstract

This research is focused on the development of models based on ensemble algorithms to predict the list price of properties for sale in Valdivia (Chile) in 2020. The research is carried out following a classic 4-stage methodology (analysis, design, development, and validation). During the analysis, data is gathered and preprocessed. In the design, attributes of interest are selected and grouped in six domains, which are combined to build 16 predictive models. Comparison metrics are selected at this point: correlation coefficient, MAE and RMSE. Construction and validation are carried out entirely using the software WEKA. A total of 34 attributes and 228 properties are considered. The dataset is split up into a subset for training and test (80%) and a subset for validation (20%). List prices are predicted using a stacking ensemble with a support vector machine as meta-learner and as base learners a linear regression, a decision tree, and an artificial neural network. In the best case, predictions and actual list prices have a correlation of 90% and a percentual MAE of 26%. In conclusion, some of the proposed models can help predict list prices. However, prediction errors are still significant.

Keywords

Machine Learning, Ensemble Algorithm, Predictive Model, Real Estate Market, List Price, Hedonic Models

1. Introduction

Real estate markets have always been controversial. Mainly due to the subjective factors involved in the negotiation of a property listed for sale. It is important to distinguish between the list price from the sales price. The list price is the amount for which a property is offered on the market. The sale price, instead, is the amount for which the transaction is done. These values are generally different, with the sales price being lower than the list price. The objective of this research is the estimation of the list price. To do it, data about properties listed for sale in Valdivia (Chile) in 2020 are obtained from the most popular real estate web sites and preprocessed. A supervised learning approach is applied to train and test predictive models based on ensemble machine learning algorithms.

Estimating the right price at which a property should be listed is crucial not only for the buyer and for the seller, but also for the loan lender involved in the transaction. In the literature, some of the most used models to estimate the price of a property are the so-called hedonic models, which are based on the principle that the price depends on the internal characteristics of the property as well as on external factors that affect it (Sopranzetti, 2015). Even though there are different approaches to develop hedonic models for real estates (Owusu-Ansah, 2013), they are not the only alternative. Models based on artificial neural networks have been also studied (Rosales and Hernández, 2021) and comparisons between different approaches have been also studied (Limsombunchai, 2004). The characteristics or attributes of a property are diverse in nature, some of them being more desirable than others. Although, there are characteristics that objectively influence the price, the selection of attributes and the complexity of the pricing models differ considerably among authors (Poeta et al., 2019). There are also investigations focused on the influence of the location on the price (Ottensmanna et al., 2008), while others propose a hierarchical or multilevel approach (Djurdjevica et al. 2008).

In this investigation, 33 attributes groped in 6 domains are considered to predict the list price. Since the final sales price can fluctuate in a wide range depending on market trends, negotiation skills and realtors' experience, it would be helpful to have a model to estimate the initial list price based in a structured method avoiding speculative factors. Therefore, it is interesting to explore the usefulness of ensemble machine learning schemes to help predict the list prices.

1.1 Objective

To apply machine learning techniques to predict the list price of properties listed for sale in Valdivia (Chile) during 2020 by means of building predictive models based on ensemble algorithms.

2. Literature Review

2.1 Machine learning

Machine learning is usually referred as a branch of artificial intelligence (AI) that uses algorithms to learn from data through experience. There are supervised, unsupervised, and reinforcement learning algorithms. In supervised learning, the training is carried out using datasets that contain the value to be predicted. In unsupervised learning, instead, the desired values or class is not known. In the reinforcement learning, on the other hand, predefined actions, parameters, and final values are used.

2.2 Cross-validation

In the context of this investigation, a dataset can be seen as a matrix made of rows and columns. While columns represent the attributes of the properties listed for sale, rows are the instances that contain all the actual values of the attributes of a single property. Thus, the dataset used in this investigation is a matrix of 228 rows (properties) by 34 columns (attributes). Since it is supervised learning task, the last column contains the desired attribute to be predicted, i.e., the list price.

Cross-validation refers to the random split up of a dataset into k folds that contain a given numbers of instances. During the model building, while k-1 folds are used for training, the left one is used for testing to assess the model's performance. Training and testing are repeated iteratively k times until all folds have been used for training and for testing (Figure 1). The objective of such iteration is to minimize the risk of the overfitting that may happen when doing a simple hold out. In the case of a cross-validation, the result of each iteration is different because the folds for training and for testing have been interchanged. The k results are finally averaged.

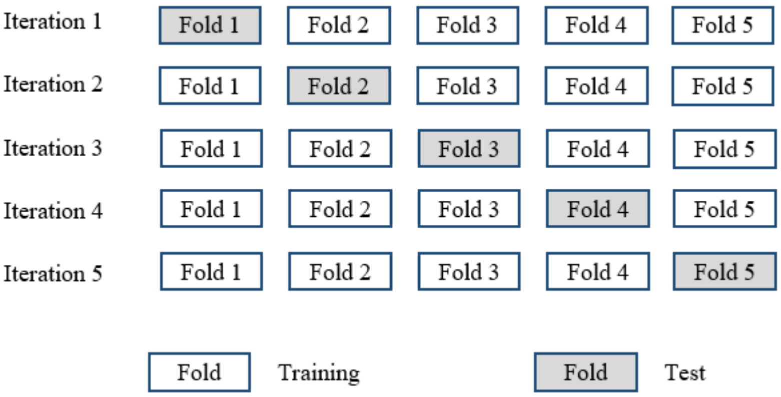
Iteration 1 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Iteration 2 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Iteration 3 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Iteration 4 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Iteration 5 Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Fold Training
Fold Test

Figure 1. Cross-validation (k=5)

2.3 Overfitting

Overfitting occurs when a model learns from the training dataset so well that it does not have a good performance when tested on an unseen dataset, i.e., the model is not able to generalize. The problem happens due to the incorporation of details from the training dataset that will be most likely not found in new data (Figure 2).

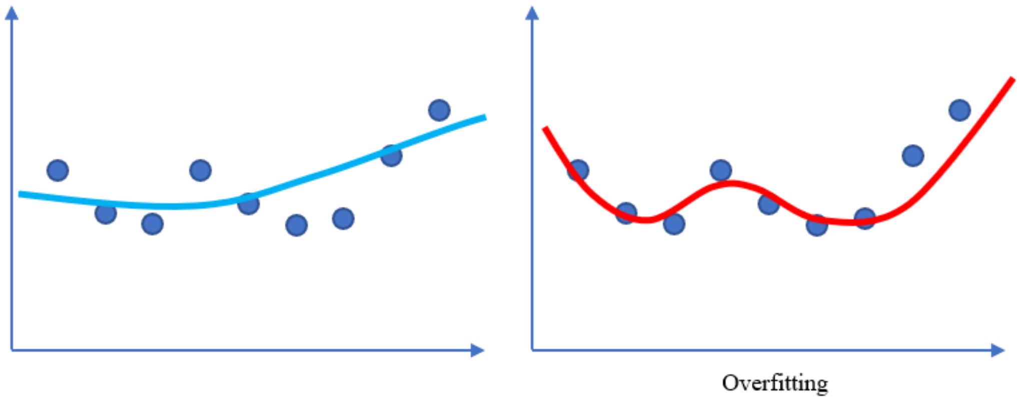

Figure 2. Overfitting

2.4 Generalization

Generalization is the ability of a trained model to classify or predict with unseen data. The objective of building predictive or classification models is to achieve a good performance with new data. The usefulness of the model depends on that. All the predictive models proposed in this investigation are evaluated on an unseen validation dataset.

2.5 Replication

Replication is repetition of an experiment under similar conditions to estimate the variability of the phenomenon under study. When using cross-validation, the k folds for training and for testing can be defined in different ways, it all depends on a specific seed number used for the partitioning. Thus, different seeds give rise to different folds and, consequently to different results. From these k results, the mean and standard deviation can be computed and analized.

2.6 Meta-learning

Meta-learning, or learning to learn, is the use of learning algorithms to learn from the prediction of other learning algorithms. The underlying idea is to combine the predictions of several machine learning algorithms to make new predictions.

2.7 Ensemble algorithms

Ensemble machine learning algorithms are multi-level structures to carry out learning tasks. In the simplest configuration there is a meta-algorithm (level 1) that leans from the predictions made by the so-called base algorithms (level 0). The ensemble usually can predict better than any of its single algorithms. The stacking is one of the most widely used ensemble schemes (Figure 3).

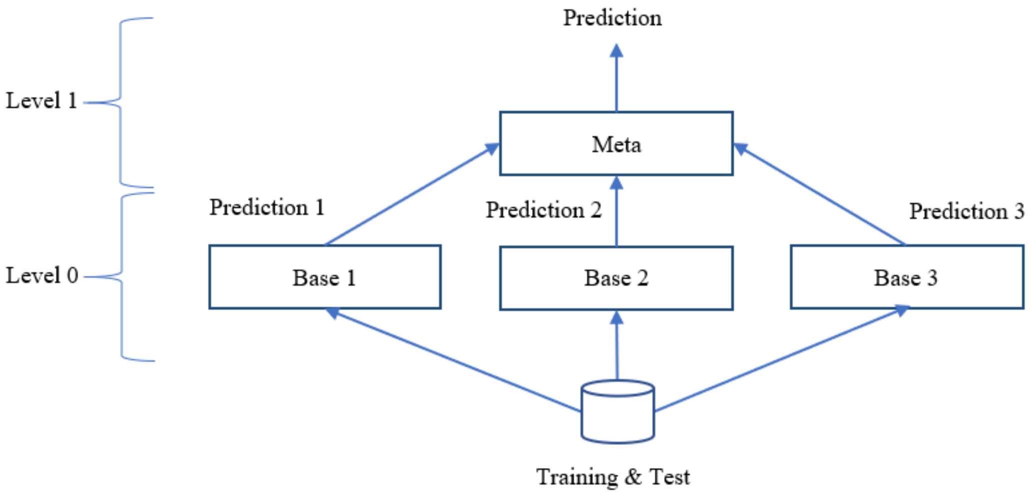

Figure 3. Ensemble algorithm

3. Methods

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

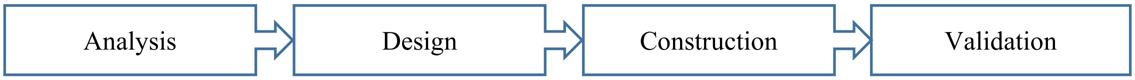

Figure 4. Four-stage model

3.1 Analysis

During the analysis, information about properties listed for sale in Valdivia (Chile) during 2020 is collected from popular real estate web sites. A total of 228 properties are considered. The scope of the investigation is defined at this point and an appropriate data processing software is selected. Attributes are grouped in several domains and for most of them, a simplified numeric scale is used (Table 1). The attribute corresponding to the list price is left as the value to be predicted. A total of 33 attributed, plus the list prices, grouped into 6 domains are finally defined for the research (Table 2).

Table 1. Domains

	Domain	**Attributes**
D.1	Property area and distribution	5
D.2	Additional features	7
D.3	Property style and condition	4
D.4	Neighborhood or district	6
D.5	Closeness to desirable amenities	7
D.6	Closeness to not desirable businesses and facilities	4

Table 2. Attributes

	Attribute	**D.1**	**D.2**	**D.3**	**D.4**	**D.5**	**D.6**
A.1	Total area m^2	✓					
A.2	Constructed area m^2	✓					
A.3	Bathrooms	✓					
A.4	Bedrooms	✓					

A.5	Stories	✓					
A.6	Front yard		✓				
A.7	Swimming pool		✓				
A.8	Parking lot/garage		✓				
A.9	Additions		✓				
A.10	Fence		✓				
A.11	Cellar		✓				
A.12	Barbecue pit		✓				
A.13	Property type			✓			
A.14	Architectural style			✓			
A.14	Actual condition			✓			
A.16	Rental property			✓			
A.17	Type of zoning				✓		
A.18	District density				✓		
A.19	Air pollution				✓		
A.20	Exclusivity				✓		
A.21	District				✓		
A.22	Accessibility				✓		
A.23	Health centers					✓	
A.24	Schools					✓	
A.25	Fire stations					✓	
A.26	Police stations					✓	
A.27	Public transportation					✓	
A.28	Supermarkets					✓	
A.29	Commercial centers					✓	
A.30	Cemetery						✓
A.31	Gas stations						✓
A.32	Industrial complex						✓
A.33	Jail house						✓

3.2 Design

The complete dataset, a matrix of 228 rows (instances) by 34 columns (33 attributes and the actual list price), is split up to create 2 subsets. The first one for training and for testing, contains 80% of the data (180 instances). The remaining 20% of data is left in a separate dataset to be used during the validation (48 instances).

To identity the relevance of each domain on the predictions, different combinations are defined to build and compare 16 predictive models (Table 3).

Table 3. Predictive models' design

	M0	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**	**M11**	**M12**	**M13**	**M14**	**M15**	**M16**
D.1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
D.2	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓				
D.3	✓	✓		✓	✓	✓		✓	✓		✓			✓			
D.4	✓	✓	✓		✓	✓			✓						✓		
D.5	✓	✓	✓	✓		✓				✓	✓					✓	
D.6	✓	✓	✓	✓	✓					✓	✓						✓

The resulting models are compared by means of the corresponding correlation coefficient, MAE (mean absolute error), and RMSE (root mean squared error) (Weijie and Yanmin, 2018).

For the purposes of this investigation, based on preliminary trial and error experiments with the training and test dataset, four machine learning algorithms are selected. Namely, support vector machine, linear regression, decision tree M5, and artificial neural network (Table 4). The ensemble scheme selected is the well-known stacking (Divina et al., 2018)

Table 4. Ensemble algorithm's configuration

Ensemble	Meta-algorithm (level 1)	Base algorithms (level 0)
Stacking	Support vector machine	Linear regression
		Decision tree M5
		Artificial neural network

3.3 Construction

All the models presented in this work are built using the software WEKA (Witten et al., 2017). Initially, models are trained and tested using a cross-validation scheme of k=10 folds with a dataset of 180 instances (Table 5).

Table 5. Models' performance with training and test data and cross-validation k=10 (180 instances)

Model	Attributes	Correlation (%)	MAE (USD)	RMSE (USD)
M0	34	89.30	40,248.32	59,686.14
M1	27	89.65	39,214.97	58,754.33
M2	30	88.96	40,430.17	60,414.38
M3	29	89.20	39,792.96	59,830.27
M4	27	89.24	40,465.59	59,682.07
M5	30	89.41	39,964.47	59,251.59
M6	13	86.54	47,150.79	66,197.18
M7	17	86.75	47,124.52	65,619.28
M8	23	89.89	39,600.13	57,957.87
M9	24	87.03	48,103.29	65,015.41
M10	28	86.63	46,612.57	65,800.09
M11	6	86.90	45,677.47	65,091.51
M12	8	43.42	90,538.40	119,904.60
M13	5	62.08	76,432.47	104,549.81
M14	7	78.05	55,884.87	82,730.94
M15	8	57.14	79,644.16	108,319.86
M16	5	49.32	85,918.65	115,900.33

Although cross-validation helps reduce the risk of overfitting, the effect of the fold partitioning remains (Powers and Atyabi, 2012). To minimize this problem, experiments are replicated with a different fold partitioning each time. For the purposes of this research 10 replications are run, which means that each model is trained and tested 100 times. The results reveal the magnitude of the standard deviation for each performance indicator (Table 6).

Table 6. Models' performance with training and test data, k=10 cross-validation and 10 replications

	Correlation (%)		MAE (USD)		RMSE (USD)	
Model	Mean	St. Deviation	Mean	St. Deviation	Mean	St. Deviation
M0	90.51	5.43	39,064.87	9,784.91	56,004.23	18,221.42
M1	90.88	5.36	38,594.15	9,616.24	55,201.46	18,131.36
M2	90.17	5.39	40,062.02	9,639.62	57,049.15	17,586.38
M3	89.94	6.34	40,018.72	10,446.42	57,460.19	18,376.68

M4	90.39	5.48	39,119.04	9,763.05	56,238.20	17,717.72
M5	90.50	5.55	39,248.64	9,664.11	55,847.12	17,964.20
M6	86.67	8.07	48,180.23	1,0833.20	64,934.85	16,239.01
M7	87.24	7.50	46,870.78	10,506.64	63,614.41	16,859.62
M8	90.62	5.49	39,181.84	9,985.66	55,489.64	17,654.04
M9	86.83	7.12	48,857.57	10,284.77	65,091.91	15,696.27
M10	87.34	7.33	46,493.95	10,287.92	64,146.12	17,497.76
M11	86.30	8.49	47,238.58	11,075.51	64,948.50	16,559.65
M12	43.15	20.20	91,270.95	17,704.49	118,417.08	24,179.70
M13	63.63	18.09	77,109.31	16,751.47	102,130.02	25,843.44
M14	79.74	8.47	56,284.54	13,278.05	80,950.95	23,225.40
M15	60.06	17.40	79,050.65	16,484.47	105,756.70	24,872.52
M16	53.11	20.67	85,260.95	16,869.07	113,869.45	24,987.46

3.4 Validation

The validation of the proposed models is carried out with the validation dataset held out during the stage of analysis. The validation dataset contains unseen 48 instances, which are totally unknown to the predictive models.

4. Data Collection

The validation results show that all models can generalize relatively well with unseen data. In all cases the percentage of correct classification is consistent with the percentage obtained with the training set. Although, both MAE and RMSE exhibit significant increments, the numbers in USD can be deceiving since they are not expressed as percentage (Table 7).

Table 7. Models' performance with validation data (unseen 48 instances)

Model	Correlation (%)	MAE (USD)	RMSE (USD)
M0	87.26	64,797.70	122,782.36
M1	87.22	65,132.39	124,635.60
M2	88.07	63,103.43	120,138.57
M3	91.17	36,583.37	54,354.25
M4	93.32	32,806.04	48,305.61
M5	87.22	64,915.02	119,948.14
M6	87.60	68,528.95	123,029.71
M7	88.92	68,174.91	121,682.87
M8	87.10	64,587.07	121,534.35
M9	87.50	67,765.20	127,417.91
M10	88.70	66,807.32	119,124.49
M11	85.64	67,423.58	122,438.59
M12	35.26	123,312.61	204,064.28
M13	69.12	100,217.50	171,563.69
M14	74.47	96,424.60	166,064.75
M15	69.90	77,459.17	112,675.15
M16	45.31	123,627.77	208,377.81

5. Results and Discussion

The results show significant differences in the performance of the models, which indicates that the attribute selection is relevant for predicting (Table 8). Even though not explained in this paper, the inclusion of the linear regression as one of the 3 base learners indicates that a correlation matrix analysis, R^2 adjusted numbers, and the p-values are being considered. In fact, even when all the attributes are added to a model (e.g., M0 with 34 attributes), not all of them are finally expressed in the resulting mathematical formulation. Although the performance of several models is similar, they are different in complexity. While some are built using the full set of attributes, others are build using just a few of them (Table 8).

Table 8. Models' size (attributes)

M0	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**	**M11**	**M12**	**M13**	**M14**	**M15**	**M16**
34	27	30	29	27	30	13	17	23	24	28	6	8	5	7	8	5

5.1 Numerical Results

Two of the most promising models are M4 and M11. Both are quite different in size, having 27 and 6 attributes respectively. The comparison between their predictions reveals correlation coefficients of 91% and 85%, and percentual MAEs of 26% and 30% respectively (Table 9).

Table 9. Actual list price vs prediction: M4 and M11 (unseen 48 instances)

		Model 4 (27 attributes, MAE 26.35%)		**Model 11 (6 attributes, MAE 30.45%)**	
Instance	**Actual list price USD**	**Prediction USD**	**Abs. Error (%)**	**Prediction USD**	**Abs. Error (%)**
1	65,616.80	138,041.67	110.38	111,917.14	70.56
2	82,677.17	111,337.89	34.67	82,513.13	0.20
3	422,080.08	369,879.57	12.37	394,174.83	6.61
4	320,328.64	314,630.91	1.78	305,769.61	4.55
5	410,774.37	440,972.62	7.35	475,388.43	15.73
6	249,343.83	209,566.18	15.95	263,968.41	5.87
7	292,931.12	258,372.52	11.80	208,427.03	28.85
8	376,178.88	415,308.68	10.40	460,537.75	22.43
9	144,356.29	240,450.60	66.57	305,826.64	111.86
10	187,976.38	238,694.92	26.98	166,788.37	11.27
11	65,616.49	86,121.79	31.25	123,691.08	88.51
12	91,863.09	82,410.93	10.29	96,052.53	4.56
13	72,178.33	59,739.81	17.23	109,237.83	51.34
14	44,641.38	61,193.87	37.08	114,315.42	156.08
15	301,987.75	262,451.75	13.09	273,333.48	9.49
16	393,700.79	301,771.06	23.35	236,175.42	40.01
17	1,205,943.10	553,488.98	54.10	597,622.90	50.44
18	157,480.32	115,585.10	26.60	156,275.24	0.77
19	546,442.97	359,418.38	34.23	408,751.32	25.20
20	102,361.58	103,242.02	0.86	94,523.61	7.66
21	89,894.77	111,005.10	23.48	67,066.48	25.39
22	446,575.80	293,643.36	34.25	231,015.00	48.27
23	113,057.17	115,982.48	2.59	111,787.79	1.12
24	741,436.43	398,443.19	46.26	393,411.02	46.94
25	170,603.67	204,182.09	19.68	97,207.43	43.02

26	306,121.12	192,154.12	37.23	213,143.46	30.37
27	194,225.72	155,485.87	19.95	156,358.04	19.50
28	419,947.51	301,868.39	28.12	264,386.17	37.04
29	248,725.76	178,482.72	28.24	202,056.08	18.76
30	105,143.16	181,925.02	73.03	144,891.54	37.80
31	118,110.24	169,024.72	43.11	137,059.27	16.04
32	86,613.85	94,514.63	9.12	102,690.38	18.56
33	81,364.60	93,613.76	15.05	84,850.14	4.28
34	419,945.57	326,913.08	22.15	360,111.63	14.25
35	112,539.36	97,288.46	13.55	116,274.63	3.32
36	244,938.73	181,006.92	26.10	230,607.75	5.85
37	85,344.22	55,879.56	34.52	135,649.54	58.94
38	90,031.57	123,711.58	37.41	89,426.66	0.67
39	170,603.67	237,080.82	38.97	289,293.36	69.57
40	85,301.84	85,725.85	0.50	82,197.27	3.64
41	282,642.91	318,936.94	12.84	312,888.76	10.70
42	95,800.12	120,859.49	26.16	123,488.94	28.90
43	65,616.80	35,660.76	45.65	100,652.35	53.39
44	91,863.52	82,835.43	9.83	140,395.92	52.83
45	111,548.23	89,170.21	20.06	102,490.75	8.12
46	209,973.75	234,222.44	11.55	258,537.62	23.13
47	640,657.27	650,110.96	1.48	643,416.04	0.43
48	118,109.69	162,746.51	37.79	199,617.46	69.01

6. Conclusion

The development of predictive models requires a good understanding of the real estate market since it is necessary to identify and select those internal characteristics and external factors that influence the price. A good selection of relevant attributes organized in domains, is useful to design combinations to build models that can be compared afterwards.In general, the use of an ensemble scheme to combine the features of different learning algorithms produces better results than the use single algorithms predicting alone. The use of cross-validation (k=10), instead of a simple hold out, helps get rid of the partitioning effect. Averaging ten results is better than having only one result. A step further is the replication of the experiment, which helps reduce the influence fold partitioning. On the other hand, having a validation dataset with unseen data is crucial to determine whether the model can generalize properly or not. Even though correlation coefficients between the actual list prices and predictions from training and test data are consistent with those from the validation data, there are significant differences when comparing the corresponding MAEs and RMSEs. This fact confirms the importance of holding out a portion of data for validation purposes. Trusting merely on metrics obtained from training and test data could be misleading.

A comparison between two models, M4 and M11, shows that not always building models with more attributes produces better predictions. While M4 is build using 27 attributes, M11 uses only 6. Being the corresponding correlation coefficients close to 91% and 85%, and the percentual MAEs close to 26% and 30%. Finally, the comparison between actual list prices and the predictions suggests that some of the proposed predictive models can predict with certain level of reliability. Although in some cases predictions are far from the actual prices, it must be understood that sellers tend to estimate biased prices that satisfy their own expectation but that might not be realistic.

References

Djurdjevica, D., Eugsterb, C., and Haaseb, R., Estimation of Hedonic Models Using a Multilevel Approach: An Application for the Swiss Rental Market, *Swiss Journal of Economics and Statistics*, vol. 144 (4), pp. 679–701, 2008.

Divina, F., Gilson, A., Gómez-Vela, F., Garcia, M., and Torres, J., Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting, *Energies 2018*, 11, 949, 2018

Limsombunchai V., House Price Prediction: Hedonic Price Model vs. Artificial Neural Network, *2004 NZARES Conference*, Blenheim, New Zealand. 2004.

Ottensmanna, J., Paytona S., and Manb, J., Urban Location and Housing Prices within a Hedonic Model Journal of Regional Analysis and Policy, vol. 38, nr. 1, pp. 19-35, 2008.

Owusu-Ansah, A., A review of hedonic pricing models in housing research, *A Compendium of International Real Estate and Construction Issues.* vol. 1, pp. 17-38, 2013

Poeta, S., Gerhardt, T., Stumpf, M., Hedonic price analysis of single-family housing. *Revista Ingeniería de Construcción, v*ol. 34, nro. 2, pp. 215-220, 2019.

Powers, D., Atyabi, A., The Problem of Cross-Validation: Averaging and Bias, Repetition and Significance. *2012 Spring World Congress on Engineering and Technology*, SCET 2012 - Proceedings. 1-5, 2012.

Sopranzetti, B., Hedonic Regression Models, *Handbook of Financial Econometrics and Statistics*. Springer, New York, 2015.

Weijie, W., Yanmin, L., Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model, *IOP Conference Series: Materials Science and Engineering*, 2018.

Witten, I., Frank, E., Hall, M., and Pal, C., *Data Mining: Practical Machine Learning Tools and Techniques*, 4th Edition, Morgan Kaufmann, Cambridge, 2017.

Biographies

Carlos Hernández is an industrial engineer, consultant, and university professor. He earned Licentiate Degree in Engineering from Universidad de La Frontera, Temuco, Chile, Master of Sciences in Computational Engineering and Doctor of Engineering from Technische Universität Braunschweig, Brunswick, Germany. He is the author of several scientific and engineering articles. He has taught lectures in 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 TU Braunschweig, Germany. His research interests include manufacturing process simulation, transportation systems simulation, supply chain design and simulation, and machine learning for finances. He is a member of IEOM.

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.