

The Cross Validation in Automated Valuation Models: A Proposal for Use

Agostino Valier(✉)

Università degli Studi di Padova, via Venezia 1, 35131 Padua, Italy
valier.agostino@gmail.com

Abstract. The appraisal of large amounts of properties is often entrusted to Automated Valuation Models (AVM). At one time, only econometric models were used for this purpose. More recently, also machine learning models are used in mass appraisal techniques.

The literature has devoted much attention to assessing the performance capabilities of these models. Verification tests first train a model on a training set, then measure the prediction error of the model on a set of data not met before: the testing set. The prediction error is measured with an accuracy indicator.

However, verification on the testing set alone may be insufficient to describe the model's performance. In addition, it may not detect the existence of model bias such as overfitting.

This research proposes the use of cross validation to provide a more complete and effective evaluation of models. Ten-fold cross validation is used within 5 models (linear regression, regression tree, random forest, nearest neighbors, multilayer perception) in the assessment of 1,400 properties in the city of Turin.

The results obtained during validation provide additional information for the evaluation of the models. This information cannot be provided by the accuracy measurement when considered alone.

Keywords: AVM · Cross validation · Real estate · Valuation

1 Introduction

Mass appraisal models are techniques for the evaluation of large sets of real estate assets, carried out using common real estate data, a unique evaluation protocol and performance verification tests [1]. Mass appraisal models are automatic calculation models, and therefore they are defined with the acronym AVM (Automated Valuation Models).

Each model expresses the function that links real estate characteristics to their prices. There are different ways in which a model can be defined. On the one hand, econometric models are based on hedonic price theory. On the other hand, machine learning models that learn directly from data [2].

Each model, once it has been defined, must be verified. The model verification phase consists of a test that provides an accuracy indicator which measures the model's error in price prediction. Accuracy tests have been very successful in scientific research. They

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12253, pp. 585–596, 2020.
https://doi.org/10.1007/978-3-030-58814-4_45

have been used both for the evaluation of proposed models and for the comparison between various models.

The accuracy indicator is a quick indicator, easy to understand even for people who are not experts in writing algorithms. However, it is often insufficient to fully describe the nature and performance capabilities of the model. The accuracy indicator therefore needs to be implemented with other information that is equally capable of being understood by non-experts.

2 Framework

All tests that measure the predictive capacity of mass appraisal models adopt the same protocol. First, the real estate dataset containing the characteristics of the properties X and their prices Y-sales prices or asking prices- is divided into two sub-sets. On the one hand, the training set with which the model is trained. It tries to find the relationship between the variables X and Y. Next, the model is tested on the remaining part of the dataset (testing set). This subset contains data with which the model has not been trained, i.e. data so far unknown to the model. The actual prices (y) contained in the testing set are then compared with the predicted prices ($\hat{y}$) formulated by the model. The smaller the difference between $\hat{y}$ and y will be, the more the model can be defined as effective [3]. The accuracy measurements of mass appraisal models are widely used in the scientific literature. Many papers have as their main objective to compare different AVMs with each other based on accuracy parameters [4]. The introduction of machine learning models has increased the number and diffusion of these papers.

The first patented Automated valuation models employed regression analysis. Actually, this type of analysis is more effective in its inferential purposes than in its predictive ability. The inferential purpose is the ability to explain the role that individual regressors (property features) play in the final price [5, 6].

Machine learning models, on the other hand, are more skilled in predictive capacity. Their main weakness is their black box character: they are not able to explain the relationship between the price of the property and its characteristics. The comparison between regression analysis and machine learning models is not only a comparison between different techniques. Many authors understand it as the comparison between human intelligence and artificial intelligence in performing the same activity: evaluating [7–10].

In the field of mass appraisal the first machine learning models to be used were artificial neural networks. The first to propose accuracy tests was Borst [11]. Afterwards, Do and Grudnitiski [12] compared artificial neural networks with multiple regression by measuring their predictive capacity. These contributions triggered a debate in the scientific community, within which there were also those who refuted the superiority of machine learning models over traditional econometric models [13]. Nguyen and Cripps [14] were credited with studying the relationship between artificial neural networks and the size of their dataset. Still nowadays, artificial neural networks are successfully tested [15, 16].

Gradually, other types of machine learning models have been tested in mass appraisal techniques. Among them, the k-nearest neighbors. Such algorithms have rarely proved effective [17, 18]. In contrast, support vector machines have been shown to provide reliable estimates [19]. Even genetic algorithms have been successfully tested by researchers [20–22].

The ensemble learning models combine several individual models within a single metamodel, which offers better performance than those performed by each model considered alone. This is the case of random forests, which are the result of the aggregation of several regression trees. Their effectiveness in predicting real estate value is widely demonstrated [23–27].

The measurement of predictive effectiveness is in fact one of the main analysis tools with which to deal with the new models that have appeared more recently in the debate on real estate valuation. However, the excessive insistence that research exerts on the results of these tests risks making accuracy the only evaluation parameter of mass appraisal models [28].

This research methodology, although very widespread, has a statistical limit. The subdivision into only two subsets can create asymmetry in the data distribution. The split between training set and testing set takes place in a random way. The division into only two groups does not ensure a uniformity of the treated sub-samples. Sampling asymmetries often generate overfitting phenomena. Overfitting occurs when the model, over-trained on the training set data, is unable to predict effectively when it is provided with the testing set data. Some techniques can be used to prevent the phenomenon, the best known is cross-validation [29, 30].

3 Case Study

This study uses as data the offer prices of 1416 residential units for the year 2013. These prices have been collected from reading sales announcements.

To implement the information contained in the dataset, data from the Oict monitoring centre (Osservatorio Immobiliare Città di Torino) were used. This monitor is managed by the City of Turin, the University (Politecnico di Torino) and the local Chamber of Commerce. It is a tool that collects annually real estate prices since 2000. The Oict divides the city into 40 microzones. The microzone and its average sales value for the year 2013 have been added as variables to the dataset.

The dataset has 13 variables, described in the table below (Table 1 and Table 2).

Table 1. Description of the variables

Variable	Translation	Description
Locali	Rooms	Number of rooms
Bagni	Bathrooms	Number of bathrooms
Balconi	Terraces	Number of terraces
Box	Parking spaces	Number of private parking spaces
Affacci	Facades	Number of facades
Cantina	Cellars	Number of cellars
Ascensore	Lift	Has lift?

(*continued*)

Table 1. (*continued*)

Variable	Translation	Description
Piano	Floor level	Floor level
Piani edificio	Floors of the building	Number of floors of the building
Microzona	Microzone	Microzone in the real estate monitor
Mq	Squared meters	Seleable square meters
Prezzi medi	Mean value	Mean value (per square meter) in the microzone
Prezzo	Price	Total offer price in euros

Table 2. Description of the dataset

Variable	Type	Missing values	Range	Mean (St. Dev)	Most frequent value
Locali	Numerical	1	1–13	3,72 (1,53)	3
Bagni	Numerical	45	1–6	1,46 (0,69)	1
Balconi	Numerical	135	0–5	1,23 (0,86)	1
Box	Numerical	450	0–2	0,31 (0,49)	0
Affacci	Numerical	278	1–6	2,1 (0,75)	2
Cantina	Numerical	278	0–3	0,70 (0,54)	1
Ascensore	Dummy	0	0 (F)/1 (T)		
Piano	Numerical	53	0–13	2,90 (2,20)	1
Piani edificio	Numerical	92	0–14	5,01 (2,23)	5
Microzona	Categorical	0	Carlo Emanuele II/Vanchiglia/Rocca/San Salvario/Dante/San Secondo/De Gasperi/Duchessa Jolanda/San Donato/Palermo/Collina/Carducci/Unità d'Italia/Santa Rita/Mirafiori/San Paolo/Pozzo Strada/Parella		
Mq	Numerical	0	24-900	108,10 (70,34)	75

(*continued*)

Table 2. (*continued*)

Variable	Type	Missing values	Range	Mean (St. Dev)	Most frequent value
Prezzi medi	Numerical	0	1.523–4.242	2516,44 (759,53)	1.956
Prezzo	Numerical	0	23.000–2.500.000	297.786,01 (262.590,50)	195.000

The missing values of the variables 'Locali', 'Bagni', 'Balconi', 'Box', 'Affacci', 'Cantina', 'Piano', 'Piani edificio' have been replaced with the most frequent value of the corresponding variable. The replacement technique, called imputation, is frequently used to deal with missing values in large datasets. It consists of replacing each missing value with the average value or the most frequent value of the variable to which the missing value belongs. After the imputation process, the dataset consists of 1416 properties. Each of them is described by 13 variables. As can be seen from Fig. 1, the correlations between variables are very weak.

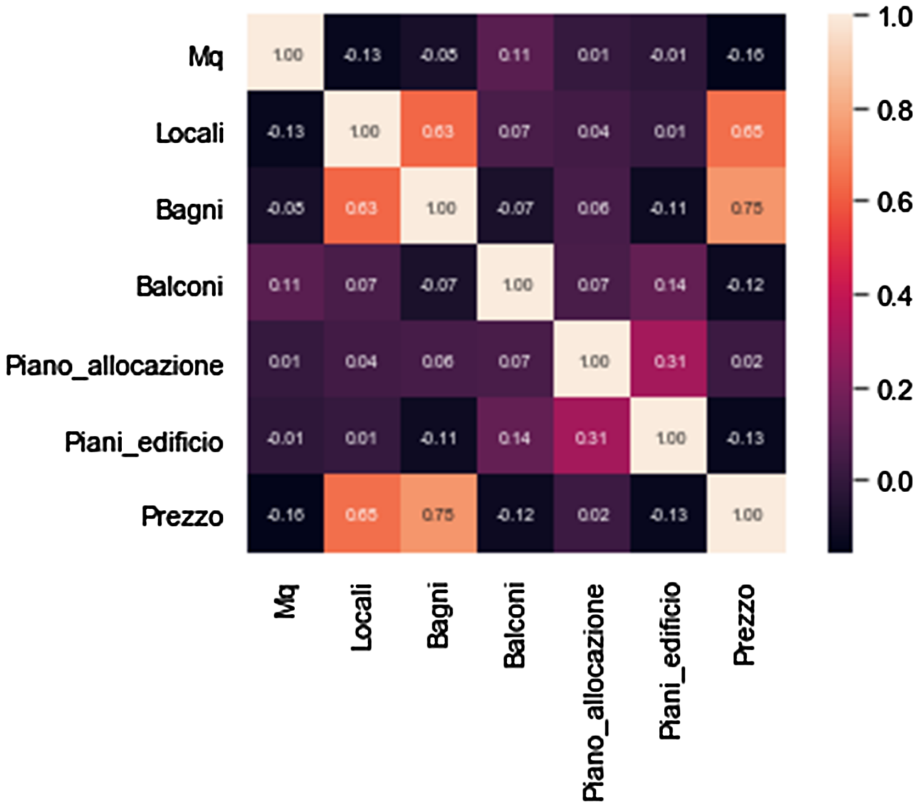

Fig. 1. Index of correlation between variables

4 Methodology and Results

The 1416 price dataset is divided into two sub-sets: the training set (75% of total data) and the testing set (25% of total data). Five models have been identified and trained. The first is a traditional econometric model, the remaining are more properly machine learning models. The models used in this research are the following:

- Linear regression
- Regression tree
- Random forest
- Nearest neighbors
- Multilayer perceptron

The models have been taken from the scikit-learn library. The GridSearchCV tool was used to detect the hyperparameters. It performs an exhaustive search by evaluating the model performance for each of the combinations in the list of values provided by the authors. It tests each combination until the optimal set is obtained.

Here the grid search has been further optimized through its use in combination with cross-validation. K-fold cross-validation is a statistical technique that consists of subdividing the training dataset into k parts of equal numerosity. At each step, the k-th part of the dataset becomes the validation dataset, while the remaining part constitutes the training dataset. Then k verifications are made by testing each time - on the data of the k-th part subtracted from the training set - the model trained on the parts of training dataset not excluded. The number k of parts with which the training set is divided is at the discretion of the authors, generally assumes the value of 5 or 10. In this research will always be used the ten-fold cross validation. This allows the model to have verification steps even before the final verification measured on the testing set. This reduces the possibility of overfitting.

The combined use of GridSearch with cross validation first identifies the best set of hyperparameters. Then it tests 10 times the algorithm (with the newly identified parameters) on the values of the training set (Table 3).

The hyperparameters found by GridSearch tool more Cross validation are the following:

Table 3. Definition of hyperparameters

Algorithm	Algorithm in Sklearn library	Hyperparameters	N° of combinations tested
Linear Regression	LinearRegression		
Regression tree	DecisionTreeRegressor	max_depth: 5	9

(*continued*)

Table 3. (*continued*)

Algorithm	Algorithm in Sklearn library	Hyperparameters	N° of combinations tested
Random forest	RandomForestRegressor	n_estimators: 9 max_depth: 8	54
Nearest Neighbors	KNeighborsRegressor	algorithm: 'brute' n_neighbors: 4	48
Multilayer Perceptron	MLPRegressor	Activation: 'identity' Alpha: 10000.0 Hidden_laye_sizes: (50,) Solver: lbfgs	960

The results are summarized in Table 4.

Table 4. Results of the cross validation (in R^2)

	Linear regression	Regression tree	Random forest	Nearest neighbors	Multilayer Perceptron
	0,693872	0,794822	0,794822	0,841647	0,905661
	0,735086	0,768665	0,684528	0,812143	0,531804
	0,679495	0,840759	0,800249	0,831772	0,879387
	0,685818	0,675845	0,628197	0,769928	0,907943
	0,628977	0,7728	0,7728	0,80649	0,79983
	0,725656	0,667922	0,70754	0,856838	0,730737
	0,734118	0,706022	0,865684	0,820216	0,914967
	0,719373	0,860305	0,865868	0,844873	0,873082
	0,713322	0,891006	0,734711	0,921169	0,450981
	0,752014	0,861159	0,860897	0,854688	0,473248
Mean	0,706773	0,78393	0,77153	0,835976	0,746764
Standard deviation	0,035833	0,0804	0,082114	0,03975	0,189925
Variance	0,00128	0,00646	0,00674	0,00158	0,03607

Then the models were tested on the testing set. They recorded the scores of Table 5.

Table 5. Results on testing set

Indicator	R^2	MAPE – Mean Average Percentage Error
Desirable outcome	Closer to 1	Minimum
Linear regression	0,638	37,70
Regression tree	0,828	24,31
Random forest	0,864	19,38
Nearest neighbors	0,820	22,15
Multilayer Perception	0,775	35,10

5 Analysis of the Results

Reading the results shows a clear superiority of machine learning models over linear regression analysis. The econometric model - here represented by linear regression analysis - records the lowest scores.

However, this research is not limited to the reading of the testing scores. The analysis of the results is divided into two parts. In the first part only the validation scores are analyzed, in the second part the validation scores are related to the testing scores.

5.1 Part I

The aim is to verify whether the models - beyond the differences in their internal structures - have a different behaviour in predicting results. For this purpose, an ANOVA single factor analysis was carried out of 50 results obtained during the cross-validation phase (Table 6).

The F value is greater than F critical value. The H_0 hypothesis, according to which the mean values recorded by all models are equal, is therefore rejected. Also the p-value, lower than the alpha value, confirms the rejection of the hypothesis H_0.

In Table 7 the ANOVA analysis is repeated, this time only on the 4 machine learning models (Regression tree, Random forest, Nearest neighbors and Multilayer Perceptron).

In this case F greatly exceeds F critical value, just as p-value exceeds alpha value. The hypothesis that the statistical behaviours of the various groups are similar to each other is therefore accepted.

These two ANOVA analyses show that the differences between traditional econometric models and machine learning models are not limited to the internal structures of the models. In fact, there are also differences in the results obtained and in their distribution.

Table 6. Anova of all five algorithms

SUMMARY				
Groups	Count	Sum	Mean	Variance
Linear regression	10	7,067731	0,706773	0,001284
Regression tree	10	7,839305	0,78393	0,006464
Random forest	10	7,715297	0,77153	0,006743
Nearest neighbors	10	8,359763	0,835976	0,00158
Multilayer perception	10	7,46764	0,746764	0,036071

ANOVA							
Source of variation	SS	df	MS	F	alpha	P - value	F crit
Between groups	0,090818	4	0,022704	2,177156	0,1	0,086813	2,074151
Within groups		45	0,010428				
total		49					

Table 7. Anova of only four machine learning algorithms

SUMMARY				
Groups	Count	Sum	Mean	Variance
Regression tree	10	78,39,305	0,78393	0,006464
Random forest	10	77,15,297	0,77153	0,006743
Nearest neighbors	10	83,59,763	0,835976	0,00158
Multilayer perception	10	7,46,764	0,746764	0,036071

ANOVA							
Source of variation	SS	df	MS	F	alpha	P - value	F crit
Between groups	0,042	3	0,014	1,112	0,1	0,356877	22,42,605
Within groups	0,457	36	0,012				
total	0,500	39					

5.2 Part II

The second part of the analysis correlates the results of cross-validation with the testing scores. The models that in the cross-validation phase recorded 10 similar values (and therefore, a low variance value) did not increase their performance once evaluated on the testing set. On the contrary, the scores decreased. The linear regression suffered a drastic drop in performance (from 0.70 to 0.63). The decrease of k nearest neighbors was more contained (from 0.83 to 0.82).

The multilayer perception algorithm records too high variance values, so it cannot be compared with the previous two. Very high values in the variation of the values indicate

a random behavior of the model, so it is difficult to make predictions. Actually, on the testing set it has not suffered a drop-in performance but an increase.

The other two models (Regression tree and Random forest) have higher variance values. This means that the values obtained in the cross-validation phase are distributed over a wider range. These models experienced an increase in predictive performance once tested on the testing set. Their final accuracy results are higher than the average of the 10 cross validation values.

6 Conclusions

The research uses the cross-validation tool to provide additional information to forecasting accuracy tests, commonly summarized in the accuracy parameter alone.

Five models (one econometric, the remaining four machine learning) were used to predict value using 1416 properties in the city of Turin. The research came to two conclusions.

The first conclusion was obtained through the use of ANOVA analysis on the results of cross-validation. The analysis shows a different behaviour between the econometric model and the machine learning models. In fact, if all 5 models are considered together the hypothesis that the mean values are equal is denied. This hypothesis is confirmed if only the 4 machine learning models are considered. The cross-validation highlights - also from the statistical point of view - the different way in which traditional models act instead of artificial intelligence models do.

The second conclusion shows that models whose validation scores have a low variance lose their predictive efficacy once verified on the testing set. The average of the values obtained in the validation phase is in fact higher than the final accuracy output. Vicecersa, models with higher variance of the validation scores, during the testing phase can obtain higher scores than the average of the results obtained in the validation phase. This is not true for models that have such a high variance that their results have a semi-aleatory behavior.

The models tested in this research are not sufficient to give a general character to these conclusions, more studies will be needed to investigate the phenomenon.

References

1. IAAO: Standard on Mass Appraisal of Real Property. International Association of Assessing Officers (2013)
2. Athey, S.: The impact of machine learning on economics. In: Agrawal, A., Gans, J., Goldfarb, A. (eds.) The Economics of Artificial Intelligence: an Agenda, University Chicago Press, pp. 507–522 (2019)
3. McCluskey, W.J., McCord, M., Davis, P.T., Haran, M., McIlhatton, D.: Prediction accuracy in mass appraisal: a comparison of modern approaches. J. Prop. Res. **30**(4), 239–265 (2013)
4. Valier, A.: Who performs better? AVMs vs hedonic models. Journal of Property Investment and Finance, article in press (2020)
5. Mangialardo, A., Micelli, E., Saccani, F.: Does sustainability affect real estate market values? Empirical evidence from the office buildings market in Milan (Italy). Sustainability **11**(1), 12 (2018)

6. Mangialardo, A., Micelli, E.: New bottom-up approaches to enhance public real estate property. In: Stanghellini, S., Morano, P., Bottero, M., Oppio, A. (eds.) Appraisal: From Theory to Practice: Results of Siev 2015, pp. 53–62. Springer, Heidelberg (2017)
7. Rosen, S.: Hedonic prices and implicit markets: product differentiation in pure competition. J. Polit. Econ. **82**(1), 34–55 (1974)
8. D'Amato, M., Kauko, T.: Advances in Automated Valuation Modeling. SSDC, vol. 86. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-49746-4_22
9. Kauko, T., D'Amato, M.: Mass appraisal methods: an international perspective for property valuers. Int. J. Strateg. Prop. Manag. **13**(4), 359–364 (2008)
10. Pérez-Rave, J.I., Correa-Morales, J.C., González-Echavarrìa, F.: A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes of real estate prices for inferential and predictive purposes. J. Prop. Res. **36**(1), 59–96 (2019)
11. Borst, R.: Artificial neural networks: the next modelling/calibration technology for the assessment community. J. Prop. Tax **10**(1), 69–94 (1991)
12. Do, A.Q., Grudnitski, G.: A neural network approach to residential property appraisal. Real Estate Appraiser **58**, 38–45 (1992)
13. Worzala, E., Lenk, M., Silva, A.: An exploration of neural networks and its application to real estate valuation. J. Real Estate Res. **10**(2), 185–201 (1995)
14. Nguyen, N., Cripps, A.: Predicting housing value: a comparison of multiple regression analysis and artificial neural networks. J. Real Estate Res. **22**(3), 313–336 (2001)
15. Núñez Tabales, J.M., Caridad y Ocerin, J.M., Rey Carmona, F.J.: Artificial neural networks for predicting real estate prices. Cuantitativos para la Economia y la Empresa **15**(1), 29–44 (2013)
16. Yacim, J.A., Boshoff, D.G.B.: Impact of artificial neural networks training algorithms on accurate prediction of property values. J. Real Estate Res. **40**(3), 375–418 (2018)
17. Isakson, H.R.: Valuation analysis of commercial real estate using the nearest neighbors appraisal technique. Growth Change **19**(2), 11–24 (1988)
18. Borde, S., Rane, A., Shende, G., Shetty, G.: Real estate investment advising using machine learning. Int. Res. J. Eng. Technol. **4**(3), 1821–1825 (2017)
19. Kontrimas, V., Verikas, A.: The mass appraisal of the real estate by computational intelligence. Appl. Soft Comput. **11**(1), 443–448 (2011)
20. Del Giudice, V., De Paola, P., Forte, F.: Using genetic algorithms for real estate appraisals. Buildings **7**(2), 31 (2017)
21. Tajani, F., Morano, P., Locurcio, M., D'Addabbo, N.: Property valuations in times of crisis. artificial neural networks and evolutionary algorithms in comparison. In: Computational Science and Its Applications - ICCSA 2015, pp. 194–209 (2015)
22. Manganelli, B., De Mare, G., Nesticò, A.: Using genetic algorithms in the housing market analysis. In: Gervasi, O., Murgante, B., Misra, S., Gavrilova, M.L., Rocha, A.M.A.C., Torre, C., Taniar, D., Apduhan, B.O. (eds.) ICCSA 2015. LNCS, vol. 9157, pp. 36–45. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21470-2_3
23. Ceh, M., Kilibarda, M., Kilibarda, M., Bajat, B.: Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS Int. J. Geo-Inf. **7**(5), 168 (2018)
24. Mullainathan, S., Spiess, J.: Machine learning: an applied econometric approach. J. Econ. Perspect. **31**(2), 87–106 (2017)
25. Kok, N., Martínez-Barbosa, C.A., Koponen, E.L.: Big data in real estate? From manual appraisal to automated valuation. J. Portf. Manag. **43**(6), 202–211 (2017)
26. Antipov, E., Pokryshevskaya, E.: Mass appraisal of residential apartments: an application of random forest for valuation and a CART-based approach for model diagnostics. Expert Syst. Appl. **39**(2), 1772–1778 (2012)

27. Huang, Y.: Predicting home value in california, united states via machine learning modeling. Stat. Optim. Inf. Comput. **7**(1), 66–74 (2019)
28. Mooya, M.M.: Market value without a market: perspectives from transaction cost theory. Urban Stud. **46**(3), 687–701 (2009)
29. Baldominos, A., Blanco, I., Moreno, A., Iturrarte, R., Bernardez, O., Afonso, C.: Identifying real estate opportunities using machine learning. Appl. Sci. **8**(11), 2321 (2018)
30. Abidoye, R.B., Chan, A.P.C., Abidoye, F.A.: Predicting property price index using artificial intelligence techniques. Int. J. Hous. Market. Anal. **8**(11), 2321 (2018)