Expert Systems With Applications 171 (2021) 114590

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain

Juan Ramón Rico-Juan[a,*,1], Paloma Taltavull de La Paz[a,b,2]

[a] University of Alicante, Campus San Vte. del Raspeig, S/N. 03080, Alicante, Spain
[b] Massachusetts Institute of Technology (MIT), 77 Massachusetts Ave, Cambridge, MA, US, Department of Urban Studies and Planning, United States

ARTICLE INFO

Keywords:
Forecasting housing prices
Hedonic tools
Machine Learning
Models' explainability

ABSTRACT

Two sets of modelling tools are used to evaluate the precision of housing-price forecasts: machine learning and hedonic regression. Evidences on the prediction capacity of a range of methods points to the superiority of the random forest as it can calculate real-estate values with an error of less than 2%. This method also ranks the attributes that are most relevant to determining housing prices. Hedonic regression models are less precise but more robust as they can identify the housing attributes that most affect the level of housing prices. This empirical exercise adds new knowledge to the literature as it investigates the capacity of the random forest to identify the three dimensions of non-linearity which, from an economic theoretical point of view, would identify the reactions of different market agents. The intention of the robustness test is to check for these non-linear relationships using hedonic regression. The quantile tools also highlight non-linearities, depending on the price levels. The results show that a combination of techniques would add information on the unobservable (non-linear) relationships between housing prices and housing attributes on the real-estate market.

1. Introduction

Analysing housing prices, including their predicted prices, is a matter of great interest in the field of the economics, in general, given the relevance the evolution of residential prices has at the macroeconomic level (ECB, 2010), and for investors and the market itself. Traditionally, there has been significant controversy among those who favour using conventional models and hedonics or repeat-sales techniques to predict housing prices (Goh et al., 2013; Hwang and Quigley, 2006; Clapham et al., 2004). Among those are researchers who show that if the spatial characteristics of dwellings are taken into account, then the level of predictability is low (Hwang and Quigley, 2004). This debate has generated the use of multiple techniques that have improved the predictive capacity of these models and have helped create more precise housing-price indices (Coulson, 2012).

The evolution of technology has allowed for the acquisition and storage of higher volumes of information. Of particular note are developments in data engineering and the use of automatic learning techniques for applications in almost all sectors of the economy. A new debate is emerging on whether these techniques, which are not based on socioeconomic behavioural models, can provide better predictions than econometric methodologies. Neural network-based applications for real-time housing-price estimations have been available for years; however, academics have rejected these methods because they considered them to be 'black boxes' whose results are not explainable (see Kauko, 2003; Zhao and Hastie, 2019 on the debate about causality and machine learning).

Data scientists have developed precise methods for forecasting socioeconomic variables by using large data sets with multiple features. An example is the well-known mass valuation technique (Kauko and d'Amato, 2009), a system which combines different machine learning tools to evaluate housing prices based on a large list of previous observations. The precision achieved in on-time valuation has lead industry and financial institutions to embrace these techniques to address the need for specific low-risk valuations in their day-to-day economic and fiscal valuations. 'Mass valuations' have been largely criticised due to their lack on clarity and critical error levels; nonetheless, this method is capable of appraising the values of a massive portfolio of houses within a short time.

Data scientists and economist analysts agree that the accuracy of ML

* Corresponding author at: Dpt of Software and computing Systems, Campus San Vte del Raspeig s/n, 03080 Alicante, Spain.
E-mail addresses: juanramonrico@ua.es (J.R. Rico-Juan), paloma@ua.es (P. Taltavull de La Paz).

[1] Department of Software and Computing Systems.

[2] Department of Applied Economics.

https://doi.org/10.1016/j.eswa.2021.114590
Received 5 October 2020; Received in revised form 29 November 2020; Accepted 7 January 2021
Available online 13 January 2021

models is not what would be desired. Data analysts' techniques ignore the laws of economics (and the explanatory models that show causality) as well as the limitations econometrics imposes on the models, leaving these systems free to make inferences from a combination of data. Nonetheless, ML techniques are capable of identifying non-linear relations between the variables that (from the economic perspective) would reflect agents' behaviour when interacting in the market. The lack of explanatory capacity of the results is one of the essential criticisms made for the use of data techniques, particularly given that their results are spurious. However, recent advances in the the explainability of ML (Lundberg and Lee, 2017) models make them particularly interesting, as their predictions provide greater accuracy and more in-depth explanations; they also treat the relationships between the variables (whether linear or not) by group or on the totality of the data.

Both groups of analysts mentioned above also agree that accurate residential price estimates require a high number of household attributes. However, according to data analysts, this also generates a high number of computations. Econometricians also point to substantial problems of collinearity and endogeneity in the models and this can either complicate the estimations or make it impossible to derive them. Further, there is no agreement on the most efficient number of attributes.

In this article, we brought the two approaches together and conducted an experiment on housing-price estimations. We used the same database to assess the accuracy of each of the main attributes and to identify the fundamental differences, advantages and disadvantages of each method.

The experiment consists of predicting residential prices based on different methodologies. From an economic perspective, a hedonic tool is used with different estimation techniques. From a data analyst perspective, several models are developed to find the best predictive results using the random forest. The explanatory capabilities are evaluated showing four error indicators and then both models are compared. One of the interesting results of the random forest method is its capacity to estimate non-linear relationships and to show, in three dimensions, how the housing features relate each other. These results are also approached by using hedonic analysis.

This article presents a comparative study of two hedonic models and a selection of machine learning models that include recent explainability processes. To this end, this article is structured as follows: in Section 2, the theoretical principles are presented; in Section 3, the model and methodology are detailed; the experimentation setup evaluates several models that are described in Section 4; a discussion and a thorough analysis of the results are presented in Section 5 and, finally, the main conclusions and some ideas for future work are summarized in Section 6.

2. Background

2.1. Theoretical principles

Housing prices are discussed extensively in the literature. This indicates how fundamental housing is to the economy and society. But this product also has considerable heterogeneity in that prices depend fundamentally on location, a factor which is not clearly observable. That is, the variable 'price' is a function of an unobservable number of influences that make its estimation complex. One part of the literature focuses on an analysis of housing prices related to this product's heterogeneity, extracting differences in price that are due to the features, location or other characteristics of the houses that are being assessed for their price. Analysis that uses hedonic models captures the relationships between the prices and the characteristics of the houses being sold in differentiated product markets. This literature is well established and supports the use of a 'quality adjusted' housing price index (Rosen, 1974; Linneman, 1980; Haurin, 1991; Peek and Wilcox, 1991; Geltner, 1993; Adair et al., 1996; Clapp, 2003). It also tests the impact of the different characteristics that are associated with real east on the level of prices and their evolution (Goodman and Thibodeau, 1995; Clapp and Giaccotto, 2002; Bourassa et al., 2011), including geographical features and their limitations (Saiz, 2010).

The literature is extensive and is mainly based on Rosen's (1974) work, where he defines the composition of a complex product based on its characteristics, allowing them to contribute to the product's final price. In Rosen's model, the property's amenities contribute to its final price. The idea is that a dwelling is a combined-attribute housing-price model in which the bundle of the dwelling's characteristics and amenities composes an 'envelope' of features of a house that is situated at any location, $z = \{z_1, z_2, \ldots, z_n\}$, with z_i measuring the amount of the ith characteristic contained in a house and z being the particular bundle of attributes that identifies the particular house. The price also represents that particular house associated them $p(z) = p(\{z_1, z_2, \ldots, z_n\})$ and the group of attributes guides both the buyers and the sellers in a specific location through all of the characteristics that equalize any differences in the set of hedonic prices. As Rosen (1974) stated: the market-clearing prices $p(z)$ are determined by the distribution of consumer tastes and producer costs (Rosen, 1974:35). Consumers continuously choose among different combinations of z, which is the function that relates houses of similar price and characteristics $p(z)$ in the absence of any changes in the envelope. This means that different consumers will 'project' their different tastes onto housing prices so that price differ for every house and every housing feature.

Thus, if the value of one particular characteristic depends on consumer behaviour, then the process of adding value for this characteristic is intrinsically determined by the willingness to pay of the buyers whose utility curve is modified by that characteristic on their entry into the market. This demand behaviour is represented in Fig. 1a, which is based on Rosen's (1974:39) definition.

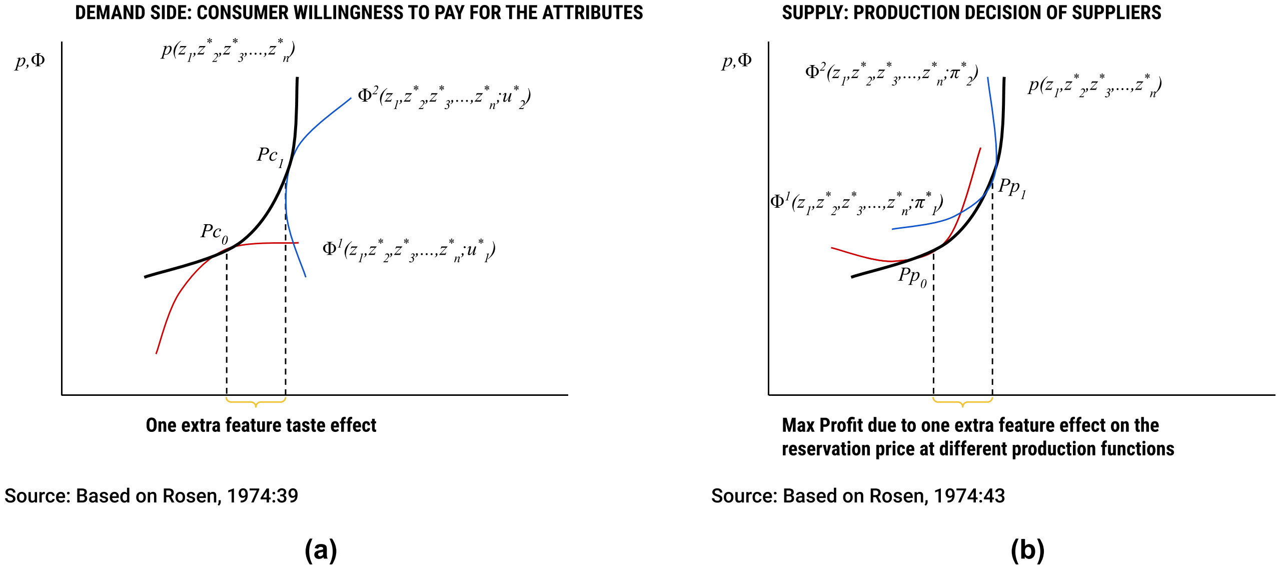
On the demand side, u_i represents the value function of every consumer at a particular location. This depends on the individual utility function and this varies according to the individual's income (y), bid function (Φ), and a combination of the selected housing characteristics ($U(y - \Phi(z_1, z_2, \ldots, z_n)) = u$). Thus, $p(z_i)$ is a function of the minimum price paid in the market for a unit bundle of characteristics with $z*$ optimum quantities of each attribute at each implicit price of z_i (Pc_0). Assuming that z_1 is the extra attribute that differs one house from another, every consumer represents a different value function.[3] Hence, a consumer whose perception is that the house with greater green attributes z is better increases the value of the green component in the envelope (bundle of characteristics). This reveals his/her willingness to pay for energy efficiency and shifts the preferred combination from the red curve to the blue, thereby determining two different fixed value functions ($u* \quad i$) and increasing the price due to the extra amount of z_1 to Pc_1.

On the supply side (Fig. 1b), the housing suppliers (sellers) exhibit an offer function ($\varphi(z_1, z_2, \ldots, z; \pi, \beta)$), which indicates the unit price they are willing to accept at a constant profit (π) for various investment designs.[4] The sellers maximize their utility at the maximum profit π, which will vary if they make a large investment to increase z_1. The increase in the latter is projected into a higher asking price (from Pp_0 to Pp_1), with the latter including the expected producer benefits that are derived from the extra investment costs.

Note that Pc_0 does not necessarily equal Pp0 and that Pc_1 is not equal to Pp_1 (nor to the attribute prices), and that the Pc's will be transformed

[3] The Φ are defined as convex functions of every combination of features, while the price function is a non-sufficient convex function representing the first order condition. *If this is sufficiently regular and convex everywhere, then higher-income consumers can purchase greater amounts of all of the relevant housing characteristics* (Rosen, 1974:40).

[4] It is assumed that the characteristics supplied in every design meet the market characteristics (the implicit prices) and achieve equilibrium at the maximum price at which one unit could be supplied at similar β coefficients.



Source: Based on Rosen, 1974:39

Source: Based on Rosen, 1974:43

(a)

(b)

Fig. 1. Rosen's 1974 demand behaviour.

into *Pp*'s after the negotiation process is complete. In addition, the above-mentioned prices (Pc_0, Pp_0, Pc_1, Pp_1) reflect the different shadow prices of the various attributes when the focus of the analysis shifts from the demand or the supply side, although both should meet the market hedonic price functions to reach equilibrium[5] in the presence of a supply that is sufficient enough to meet demand.

Some authors report having econometric issues with the hedonic models, as such they assert that these models provide limited accuracy in the estimation of house prices (Goodman and Thibodeau, 1995; Goodman and Thibodeau, 2003). In fact, hybrid models have been developed in order to avoid price underestimations and other errors (Case et al., 2006) but they are nevertheless based in the conceptual framework of hedonics. Hedonic methods' lack of capacity to capture the full behaviour of house prices is an indication that these prices play a role in internalising this market's dynamic evolution and also serve as an indicator for other purposes (Case et al., 2005).

Housing prices have been much debated from other perspectives. Adding complexity to Rosen's idea, there is a large consensus that the heterogeneous features of houses also change depending not only on consumer's expectations and wealth but also on the time period and market shocks. This is why part of the literature tends to take a long-term view of housing prices and the rationale behind them, despite the fact they are recognized as responses to short-term demand, with construction activity or vacancies also affecting how prices respond. All of these factors introduce asymmetric adjustments that has been identified as speculative bubbles (Stiglitz, 1993; Mikhed and Zemčík, 2009; Abreu and Brunnermeier, 2002; Kim and Suh, 1993). The conventional explanatory models for the behaviour of residential prices in housing economics follow either the life-cycle model, in which prices represent the long-term market equilibrium component, or they concentrate on the explanatory factors related to the price composition on the supply side (Meen, 2012; Dougherty and Van Order, 1982; Muellbauer and Murphy, 1997), including the process of financial liberalisation that promotes the growth of residential prices across most Western countries (Ortalo-Magne and Rady, 2006) and increases credit (Allen and Gale, 1998) until there is a financial crisis.

In residential markets it is argued that heterogeneity and lack of information are typical characteristics that make it more complicated to distinguish changes in house prices (Case and Shiller, 2003) when expectations of future prices rise and cause current prices to become temporarily high. Research during the 2000s shows that acceleration and temporariness would be considered a normal reaction of the residential market, adding complexity to the analysis and understanding of housing prices. In summary, (Riddel, 1999) we propose three components that define market prices for existing homes. These are the conventional component, which is based on economic and demographic variables; the feedback component, which considers changes in prices that stem from information about relationships with previous appreciations in housing prices and which introduces an error-correction mechanism that explains any price changes; and the expected component, which is driven by the prices that were observed in the previous period and which are then corrected by the forecast error that produces a systematic price appreciation beyond that motivated by economic fundamentals. This latter component should be one which can be captured by changes in shadow prices as defined according to Rosen (1974).

In addition to the above, the literature on housing prices also maintains that people's expectation of an increase in housing prices are formed when they see a growth in capital gains (Dougherty and Van Order, 1982; Poterba, 1984; Poterba et al., 1991). When capital gains grow at normal rates, this guarantees balanced market growth, but the evidence confirms that housing prices grow at rates above or below normal, depending on the period (for instance, in Black et al., 2006). Housing prices to not follow a clear time pattern which could be predicted as having a permanent effect (Meen and Andrew, 2004) over the long term; they also shows periodic reversions back to fundamental values, taking decades to recover (Mikhed and Zemčík, 2009). An increasing number of empirical papers provide evidence that housing prices ripple across regions and this is mostly because the economic factors that influence these prices differ (Meen, 1999 and others).

The last component of housing prices is the spatial influence. Other than location (which determines differences in price levels), the literature demonstrates the existence of spatial autocorrelation among housing prices in the local markets (Anselin, 2013; Taltavull de La Paz et al., 2017), which also influences the growth of housing prices.

The above summary of the literature suggests that any analysis of housing prices from an economic perspective will miss several nonlinearities that are not considered in the relationships between the

[5] Rosen (1974:44) points out, *the* $p(z)$ *represents a joint envelope of a family of value functions and another family of offer functions.*

characteristics that influence housing heterogeneity. Changes in expectations are also led by general economic conditions. The current literature has not yet dealt with the large number of variabilities which could allow us to precisely evaluate and forecast housing prices. Some of the research introduces these variabilities by using non-linear hedonic models as measurement tools and by controlling for any endogenous relationships among these variables (list literature).

3. The model and methodology

The hedonic framework we described above is the model that is conventionally used in housing economics to evaluate the role of preferences/willingness to pay for a dwelling's characteristics. The shadow prices of the relevant housing characteristics are estimated using a conventional hedonic model in which every component will reveal the bidders' preferences for each characteristic, as in Section 4.2.2

$$LPh_{it} = \alpha_{it} + \sum_{t=1}^{T} \sum_{k=1}^{n} \beta_{kt} X_{kit} + \in_{it}, \tag{1}$$

where *LPh* is the *log* of the housing asking prices (the supplier's listed price); *X* is a matrix of the housing characteristics, including the number of bedrooms, family rooms, and bathrooms, whether there is a garage, the type of house, its age, quality, location and other characteristics. α, β are the parameters to be estimated and $\in$ is an error term.

Data analyst use a methodology that is quite different from the hedonic framework. The ML-oriented approach would include the same input parameters as those in the model, so it can relate the same input features, *X*, to the variable to be predicted, *LPh*, as in Eq. 2. These types of numerical predictions are known as regression models. Since several families of algorithms can be used to perform these calculations, some would be tested by measuring the quality of their final prediction and then selecting the best one. It should be noted that these algorithms are capable of learning both linear and non-linear relationships between *X* and *LPh*. Once the model is chosen and trained with the data, explainability techniques would be used to obtain, for example, the relevance of the features at a global level by group or by sample, as well as the relationship between the input variables and their contribution to the value of the prediction.

$$LPh_{it} = model_{predict}(X_{it}) \tag{2}$$

4. Experimentation

In this section, the dataset and the models are described, as well as their results in two steps: first, the machine learning models are explained and, second, the conventional quantile and hedonic models are used to model housing prices.

4.1. Dataset

The dataset in this exercise consists of data on Alicante province (Spain) for the period 2004–2012 and is detailed in McGreal and Taltavull de La Paz (2012). Due to the large size of this dataset (1,124,502), a random selection from the full database containing 30% of the observations was chosen to run the exercise. This ensured the efficiency of the big-data processes (in terms of the computer time required for processing) for a total of 392,412 observations of individual dwellings on the market. The database was obtained from a big valuation company (Tabimed, which was the fourth largest in Spain before 2012; this company is now defunct) and provides information only on the comparable characteristics that are used for this exercise. The data are geolocalised from 2008 (figure 2), which reduces our dataset to around 56,000 observations (for five years). The data, then, refers to individual dwellings being sold on the market, showing their asking prices. The number of characteristics in this database is 52, which includes houses, buildings and neighbourhoods; other features are also available according to the geolocation and the process used to assign this extra information in the company's database. Details are shown in Table 1, which provides the main statistical features and in Fig. 3, which shows the number of houses and the average price per period.Fig. 4.

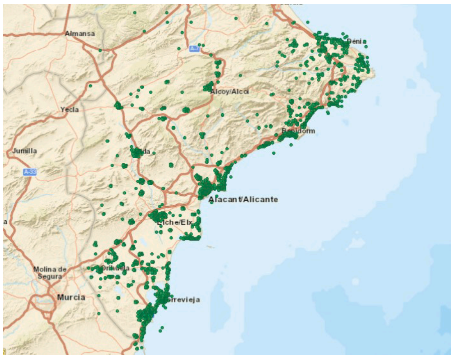


Fig. 2. Geolocalized data on individual dwelling sales in the province of Alicante (Spain) during the period 2008–2012.

4.2. Machine learning modelling

Recent publications apply ML algorithms to predict housing prices in interesting settings (Park and Bae, 2015 or Yoo et al., 2012). In this article, we test a wide range of modern and classical (linear regression) ML algorithms in order to predict the house prices that are offered during periods when prices show a high variance in the touristic province of Alicante, Spain.

Testing by using a varied number of families of algorithms helps obtain the best model(s) with which to make good predictions; it also helps to more accurately explain the relationship between the variables (input or output). In addition, this study uses ML models with advanced explainability so as to perform more accurate analyses.

4.2.1. Algorithms

To explore different alternatives that are based on machine learning, we chose algorithms that belong to different families of categories (neighbourhood, decision trees, neural networks and linear regression) in order to find the ones that provide better results for the dataset we are studying.

The algorithms considered are presented as follows, along with their brief descriptions:

- *Nearest Neighbours* (Cover and Hart, 1967): This algorithm computes a prediction value that is based on the *k* (parameter), which provides the closest examples of the training set. This model interpolates the final predictions based on the 'neighbours' proximity according to the Euclidean distance. In our case, the parameter *k* was fixed to 1, 3, 5, and 7.
- *Decision tree* (Breiman, 2017): This model predicts the value of a sample by learning simple decision rules, hierarchically. The tree is constructed from the training data and considers only one feature per rule.
- *Random Forest* (Breiman, 2001): This model builds multiple decision trees in order to combine all of the predictions for more robust behaviour. Its main parameter is the number of trees used to calculate its predictions. In our case this has been set to 100, 200, 300, 400 and 500.
- *AdaBoost (Adaptive Boosting)* (Freund and Schapire, 1997): This algorithm builds multiple linear regressors. The final decision for a test sample is taken into account according to all of the predictions, each of which is weighted by a confidence value that is learned during the training process.

Table 1
Statistical descriptors of the dataset.

Feature	Mean	Std
Log price	11.52	0.61
Year	2003.14	3.87
Postal code	3568.50	4004.16
Age of neighbourhood	15.37	15.42
Urban type	2.45	0.87
Population	90467.87	385807.99
Economic activity	3.18	1.27
Population growth	1.73	0.47
Urban Rural	2.99	0.09
Housing use	1.53	0.86
Income level	4.28	0.71
Population density	2.52	0.52
Population development	3.39	0.83
Road quality	2.93	0.34
Water source type	5.53	0.51
Water quality	2.53	0.50
Sewer type	6.52	0.63
Sewer quality	2.55	0.54
Lighting system type	3.97	0.24
Lighting system quality	2.93	0.43
Retail facilities quality	4.20	0.92
School facilities quality	3.84	0.83
Religious facilities quality	3.84	0.80
Leisure facilities quality	3.86	0.76
Sports facilities quality	3.84	0.80
Health facilities quality	3.84	0.80
Bus	4.04	0.83
Train	1.15	0.98
Underground	0.01	0.14
Housing type	1.67	0.91
Number of dweeling in building	19.50	29.83
Number of lifts	0.89	0.95
Age	7.77	10.48
Retail facilities quality in neighbourhood	4.14	1.19
Income in building	4.28	0.67
Population density neighbourhood	2.53	0.52
Location	2.89	1.61
Orientation	4.89	2.36
Views	2.52	0.81
Construction quality	3.98	0.74
Housing size (m2)	102.58	35.51
Outdoor living space	4.18	15.82
Urbanization quality	0.74	1.31
Type of housing use	1.24	0.42
Number of floors	6.71	3.89
Exterior rooms	3.03	1.45
Number of rooms	6.62	1.84
Number of bedrooms	2.78	0.85
Number of bathrooms	1.66	0.55
Latitude	38.36	1.41
Longitude	−0.53	0.61
Provincial code	3436.64	1735.11
Month	6.36	3.44

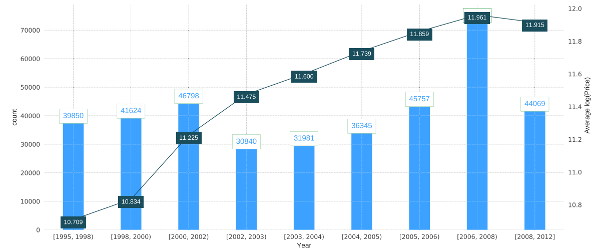


Fig. 3. This figure shows the number of houses analysed per period and their corresponding average price.

- *XGBboost)* (Chen and Guestrin, 2016): This is an extension to AdaBoost, where optimization is performed using derivable cost functions and a gradient descent (as in neural networks) is used to find the best parameters for the problem.

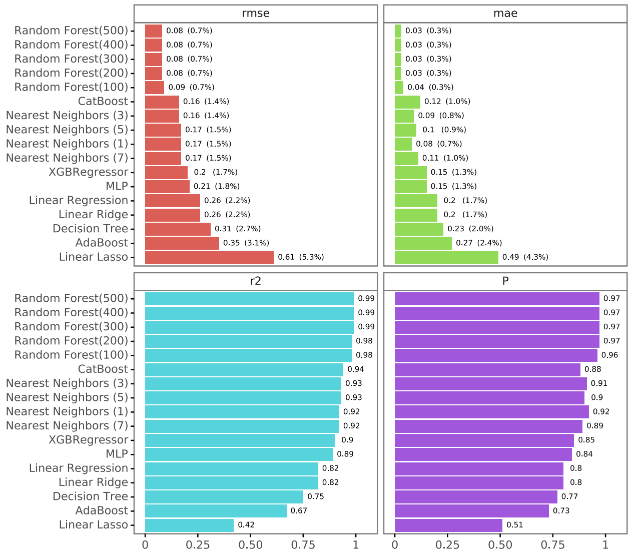


Fig. 4. Average results of 10-CV sorted by the RMSE using different metrics. RMSE, %RMSE, MAE and %MAE; the lower value is better, while for accuracy (P) and R^2 the higher value is better.

- *CatBoost* (Dorogush et al., 2018): This is another implementation of a method that is based on gradient boosting over decision trees that are similar to the previous ones that had performed well in open challenges.
- *Neural Network (Multilayer Perceptron)* (Hinton, 1990): This is the traditional neural network where all of the layers are fully connected to each other.
- *Linear regression* (Weisberg, 2005): This is a classic linear regression model that is based on linear relationships between the input features; it assumes the features are independent and optimized with a least squares approach.
- *Linear Ridge* (Hoerl and Kennard, 1970): This is a linear regression that optimizes a function using both components at the same time: a loss based on the least squares and another loss based on regularization (l2-norm).
- *Linear Lasso* (Tibshirani, 1996): This is the acronym for the model called least absolute shrinkage and selection operator. This method, similarly to the previous one, performs both variable selection and regularization in order to enhance the prediction accuracy.

4.2.2. Results

The metrics used to evaluate the models are those commonly used for this purpose. They are based on the measurement of the prediction errors or the root mean squared error (*RMSE*) (Eq. 3) is a percentage of *RMSE* with respect to the average of the true values, (Eq. 4) is the mean absolute error (*MAE*), (Eq. 5) is the percentage of *MAE* with respect to the average of the true values, (Eq. 6) is the degree of model's explanatory accuracy (P), (Eq. 7) is the R^2 score (Eq. 8). These measures are used in other articles on similar topics such as Hu et al. (2019), where y and $\hat{y}$ are vectors with true and predicted values, respectively.

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \qquad (3)$$

$$\%RMSE(y,\hat{y}) = \frac{\sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}}{|\overline{y}|} \qquad (4)$$

$$MAE(y,\hat{y}) = \frac{1}{n}\sum|y_i - \hat{y}_i| \qquad (5)$$

$$\%MAE(y,\widehat{y}) = \frac{\frac{1}{n}\sum|y_i - \widehat{y}_i|}{|\overline{y}|} \tag{6}$$

$$P(y,\widehat{y}) = \frac{\sum\left(1 - \left|\frac{y_i - \widehat{y}_i}{y_i}\right|\right)}{|y|} \tag{7}$$

$$R^2(y,\widehat{y}) = 1 - \frac{\sum(y_i - \widehat{y}_i)^2}{\sum(y_i - \overline{y})^2} \tag{8}$$

In spite of the above results, most of the reported values fall into narrow ranges; this might mean questioning the statistical significance of the differences between the algorithms considered. To verify these results, we use the Wilcoxon paired test (Wilcoxon, 1945) to determine which results are significantly better with a 95% confidence, which is the commonly used value.

Figure 5 shows a comparison of statistical significance between algorithmic pairs according to the different metrics considered in this study. It can be observed as in all the metrics that: i) the random forest algorithm obtains the best results; and ii) the results for the random forest using the parameters of 200, 300, 400 and 500 are significantly equivalent. For these reasons the random forest (200) is chosen as the best model because it is the least costly in time and space of these four best options.

4.2.3. Machine learning explainability model

Depending on how predictive models are built in machine learning, the most transparent and easy way to interpret algorithms are those based on neighbourhood criteria (e.g. kNN) or on the construction of a single decision tree (e.g. Decision Tree). Those based on multiple decision trees (e.g. Random Forest, XGBoost and CatBoost) or on artificial neural networks (e.g. Multilayer Perceptron) are more difficult to interpret (so-called black boxes) but they also tend to achieve the best results. All these models are able to learn non-linear relationships between incoming and outgoing variables which makes them more accurate in the prediction to solve real problems while classical multiple linear regressions (e.g. Linear regression, Rigde and Lasso) are easy to calculate and interpret but assume in their models only linear relationships.

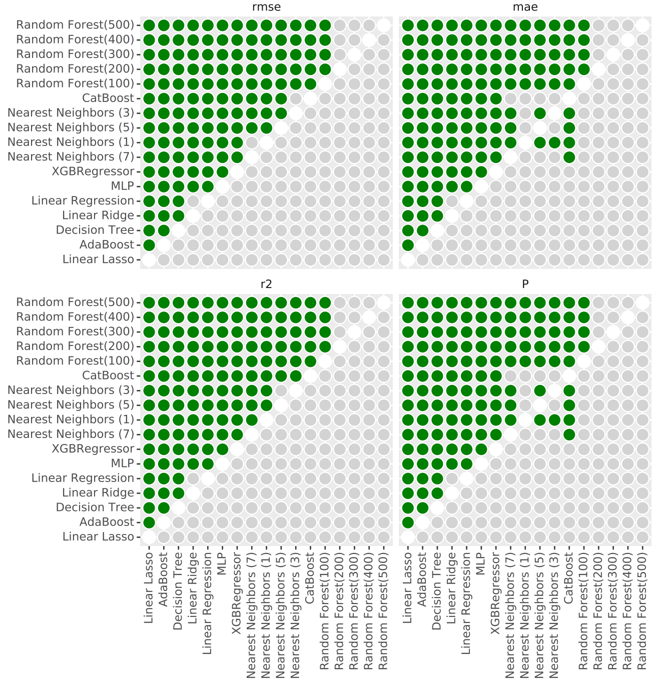


Fig. 5. Wilcoxon paired significance test. The green bullets show whether the 10-CV results of the algorithm in the row are significantly better than the algorithm in the columns for each metric used.

In general, the best results are obtained with the models we have called 'black boxes' and to try to explain their predictions in a coherent manner there are two approaches: the first is based on performing permutations (Breiman, 2001) on the value of each individual input variable and comparing the variability in the predictions. This allows us to analyse the importance of the input variables in a model that is already trained; the second approach is based on building a new linear model that explains the complex model.

The most advanced algorithm in the second approach is based on the Shapley (Roth, 1988) values that provide a solution that consists of equitably distributing profits and costs among several collaborators. This method is usually used in situations where each collaborator contributes in an unequal way. In essence, a Shapley value represents a collaborator's average expected marginal contribution after considering all possible combinations. Additionally, this method guarantees accuracy and consistency. Although, most ML models use non-linear relationships between the input variables internally, this new model's approach can be used to explain its behaviour (Fig. 6) because it uses a locally applied linear approach.

Recent advances in the approach are explained by Lundberg and Lee (2017), Lundberg and Lee (2017), who allow a unified approach to explaining the predictions made by any machine learning model. An example of this type of tool is SHAP (SHapley Additive exPlanations) (Lundberg, 2019), which connects game theory and local or global explanations by uniting several methods.

4.2.4. Importance of a model's features

The model features and their importance directly depend on the model used. In our case, we use the random forest (200) method, according to the results obtained in Section 4.2.2.

In this trained model, the Shapley values are computed individually and the absolute values for each input variable are accumulated to determine its importance. The features are then ranked in order of importance. These are presented in Table 2.

4.3. Hedonic models

Model (1) is estimated using the variables defined in the previous random forest (200), which is the best prediction model obtained. Note that some of the variables have no economic meaning. For instance, using provincial (municipal) codes, postal codes, and latitudes and longitudes as regressors makes no sense from an economic perspective even though they capture the spatial differences in housing prices. However, we do include these in the first model for comparison purposes.

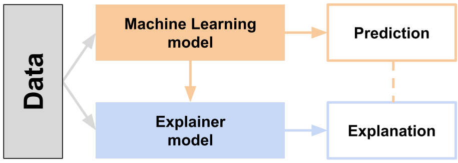


Fig. 6. General scheme of the explainability of machine learning models.

Table 3 provides the results of the non-linear hedonic model. This model includes housing features that are ordered according their relevance to explaining the variations in housing prices (absolute values of the standardised estimated coefficients are used for this purpose). The

Table 2
Ranking features by their importance in the random forest (200) model by accumulating the individual absolute Shapley values.

Rank	Feature	Importance	Rank	Feature	Importance
1	Year	3564.5	27	Population development	36.7
2	Housing size (m2)	1464.9	28	Population density	36.6
3	Income in building	877.0	29	Number of rooms	25.0
4	Age	700.9	30	Retail facilities quality	25.0
5	Number of bathrooms	678.3	31	Exterior rooms	24.0
6	Location	648.8	32	Type of housing use	22.9
7	Housing type	586.6	33	Latitude	21.7
8	Number of lifts	514.3	34	Train	17.6
9	Views	256.5	35	BUS	14.6
10	Economic activity	230.1	36	Urban_type	10.9
11	Construction quality	165.7	37	Number of bedrooms	10.0
12	Population	149.9	38	Number of floors	8.8
13	Urbanization quality	147.5	39	Health facilities quality	5.5
14	Number of dwellings in building	106.0	40	Road quality	5.3
15	Outdoor living space	105.5	41	School facilities quality	4.7
16	Retail facilities quality neighbourhood	99.5	42	Sports facilities quality	4.7
17	Orientation	92.2	43	Leisure facilities quality	4.6
18	Population growth	88.4	44	Lighting system quality	4.0
19	Housing use	79.9	45	Water quality	3.3
20	Postal code	78.7	46	Sewer quality	2.4
21	Provincial code	73.1	47	Religious facilities quality	2.3
22	Income level	60.1	48	sewer type	1.2
23	Population density neighbourhood	58.3	49	Lighting system type	0.8
24	Age of neighbourhood	56.7	50	Water source type	0.7
25	Longitude	55.9	51	Urban rural	0.3
26	Month	49.4	52	Underground	0.1

model is estimated both with and without the constant variable, the latter in order to approach the machine learning method.[6] The conventional hedonic Eq. 1 explains 80.8% of the variations in the dependent variable with an error of 17.05%. The goodness of fit is quite large compared to the other hedonic models' results for housing prices. The level of collinearity is very low, with two variables (ROAD QUALITY AND LIGHTING SYSTEM) being highly correlated to variables which should be deleted as they capture the influence of other components in the model, thereby causing a high level of bias in the estimated parameter.[7],[8]

Regarding the results from the machine Learning approaches, the non-linearities identified among the various covariates suggest that the parameters for each attribute obtained from the hedonic model could not be fixed and may have had different effects, depending on the level of the housing prices. In contrast with the random forest method's ability to show three-dimensional relationships, the hedonic model that is calculated under quantile regression methods provides estimations along two dimensions; that is, the hedonic model quantifies the different effects of the covariates (the parameter values) on housing prices in different ranges of their distribution. This is very relevant in housing markets as it can show how an increase, for example, in income in a low-income household can cause the housing price to rise but to a lesser degree than it would for a high-income household (see Fig. 7 second panel). In another example, there would be a low effect of an increase in the housing size (m2) among those houses within the low- and medium-price ranges, compared to the large effect there would be on houses within higher price ranges.[9]

In this paper, 10 quantiles have been calculated for each 10% of the housing-price distribution; that is, the effect of every covariate is calculated for the prices falling in the first 10% of the price distribution, and the second corresponds to 20% of the price distribution and so on up to 90%. The full regression corresponds with the first estimation. The results appear in B. The quantile hedonic regressions show a lower predictive power than the least squares hedonic, at around 58–59%; however, it has provided richer detail about how the covariates build housing prices as well as about the non-linearities in some of the attributes.

As these results demonstrate, the effects of the covariates vary across the price distribution, with different impacts. This supports the hypothesis that housing attributes have different effects on housing prices, depending on the price level and they indirectly capture several non-linearities that are associated to these attributes. The following figures represent variations in the covariate effects by quantile as they affect housing prices. Some attributes are selected to find evidence that is similar to those in the machine learning estimation:

The sub-figures in Fig. 7 represent how population, income, age and size affect housing prices in the different quantiles. In the first case, the effect of population remains unchanged for the first price ranges until they reach 60% of the price level. Since in the sixth quartile, population has an increasing effect on prices, this contributes to the idea that housing prices strongly increase in cities with larger populations.

[6] In econometrics, it is accepted that the constant is a key component in regression models as it captures the true conditional mean of the dependent variable. From a statistical perspective, ignoring the constant means that the other parameters are 'forced to cross the origin' when they are estimated and this results in different values and potential bias. The results from both estimations differ, with the first model showing that the dwelling features have a more accurate and widely spread influence on their prices. The estimation without a constant is available upon request.

[7] The other variables show soft collinear relationships which should be tested before accepting the final functional form of the hedonic model.

[8] A way to improve the forecast capacity in hedonic models would be to include the residuals in a second-step estimation which would allow us to predict larger parts of the dependent variables. As this method will emphasise problems of collinearity (and endogeneity), it has been fully rejected in the economic analysis.

[9] In technical terms, this means that the conditional distribution of the dependent variables related to the covariates may be asymmetric, either exhibiting unusual skewed tails or not having single modes. In the quantile regression, the analysis concerns the distribution rather than the conditional mean captures the different distributions among the quartiles in the data. The quantile regression estimator is an asymptotically normally distributed, semi-parametric method and as it is based on a median regression, it is more robust to outliers than the least squares regression that is used in conventional hedonic models. This technique allows us to consider the impact of a covariate on the entire distribution of y, and not merely its conditional mean.

Table 3
Hedonic regression (with constant). Ordered housing feature ranking by relevance on price determination. * The order follows the standardized beta from the largest to the lowest absolute value, *** p-value < 0.01, ** p-value < 0.05.

Feature	Non-standard B	St error		Collinearity test VIF
(Constante)	130.7274	1.17	***	0.00
Housing size M2	0.0071	0.00	***	2.81
Location	−0.0680	0.00	***	2.45
Year	−0.0601	0.00	***	1.32
Income building	0.1022	0.00	***	2.12
Income level	0.0936	0.00	***	2.04
Number bathrooms	0.0851	0.00	***	2.17
Age of neighbourhood	−0.0040	0.00	***	1.97
Economic activity	0.0248	0.00	***	1.64
Road quality	0.0828	0.02	***	131.09
Number of lifts	0.0346	0.00	***	1.94
Religious facility	−0.0734	0.00	***	2.85
Views	0.0290	0.00	***	1.67
Age	−0.0018	0.00	***	1.41
Type housing use	0.0333	0.00	***	3.68
Month	−0.0066	0.00	***	1.20
Housing use	0.0564	0.00	***	3.94
Number of dwellings in building	−0.0006	0.00	***	1.82
Population	0.0000	0.00	***	3.42
Exterior rooms	0.0121	0.00	***	1.30
Health facilities q	0.0375	0.01	***	7.72
Outdoor living space	0.0009	0.00	***	1.11
Population develop	−0.0183	0.00	***	1.58
Sports facilities q	0.0340	0.00	***	4.12
Retail facilities q neighbour	0.0124	0.00	***	1.90
Housing type t	0.0122	0.00	***	3.71
Construction q	0.0131	0.00	***	1.18
Urban type	−0.0127	0.00	***	3.82
Train	0.0110	0.00	***	1.91
School facilities q	−0.0234	0.00	***	5.60
Urbanization q	−0.0069	0.00	***	3.48
Longitude	0.0143	0.00	***	1.14
Lighting system quality	−0.0220	0.02		130.60
Population growth	−0.0198	0.00	***	1.50
Number of rooms	−0.0046	0.00	***	2.41
Population density	0.0139	0.00	***	2.85
Orientation	0.0025	0.00	***	1.10
Number of floors	0.0015	0.00	***	1.51
Postal code	0.0000	0.00	***	1.18
Retail facilities q	0.0048	0.00	***	2.12
Latitude	−0.0031	0.00	***	1.03
Population density neighbour	−0.0088	0.00	***	2.84
Number of bedrooms	0.0053	0.00	***	2.59
Urban rural	0.0467	0.02	***	1.15
Sewer quality	0.0616	0.03	**	1.31
Lighting system type	−0.0185	0.01	**	1.41
Bus	0.0014	0.00		1.27
Underground	−0.0075	0.01		1.00
Excluded				
Sewer type				
Leisure quality				
Goodness of fit				
Adjusted R^2	0.8088			
Standard error	0.1705			

Similar effects are seen for housing size (last chart in the Fig. 7), with almost constant effects on prices between the second and 8^{th} quartiles, but large effects for both smaller units and larger units. The different effects on prices are shown in the large queues in the distribution.

The literature notes that income is one of the determinants of housing prices. Its effect on price increases is minimal at most price levels until quartile 8. This suggests that any increase in the income level has less effect on lower-priced housing, with similar effects in a large range of prices (4th to 8th quartile of prices) and strong effects on more expensive houses. Similar interpretations are shown on Fig. 9.

Age has a great effect on lower-priced houses, compared to higher-priced domiciles. The diminishing prices that are due to age show strong effects on lower-priced units and this is possibly associated to their quality and location. On the contrary, the negative effect is much lower on the other side of the price distribution; that is, for older houses in high-priced locations, the effect of an additional year of a house's age is negative. This interpretation is different from the one in Fig. 11, which determines that the non-linearity of the age effect by itself is associated with the period during which the house is listed on the market. In quantile regressions, the negative parameter suggests a similar idea and the differences in the 'negativity' of the parameters across quantiles suggests non-linear effects, depending on the price level, but no conclusion can be inferred in relation to the time at which a house is placed on the market.

The random forest method's ability to find non-linearities in the data is superior to that of the hedonic models; however, the latter come closer to the economic logic of understanding why housing prices rise.

In the hedonics, the models identify that the errors contain related information ($DW = 0.6$), producing inefficient parameters. With these results, the estimated hedonic should be re-calculated in order to fit the basic assumptions and to obtain results that provide a better goodness of fit, so as to explain the higher part of the changes in the dependent variable.

Comparing the hedonic models with the random forest(200),in Table4 the ranking of the features that affect housing price changes seems to be closer. As can be seen, the first six features in terms of their relevance fully coincide with the association between housing prices and the economic moment (captured by the variable YEAR), the location, size (in m^2) and the income earned from the building, which are, for the most part, the main determinants of housing prices. The random forest also finds a strong relationship between AGE and LOCATION, while the hedonic model identifies the relevance of the AGE and NEIGHBOURHOOD independently. Interestingly, the neighbourhood's features appear far from the first level of influence in both models.Table 5.

While the hedonic method provides a 'measure' of the influence on housing prices in the form of a parameter, the random forest does not. Using the latter for only a few variables allows us to provide a high quality prediction, while the hedonic method is not that precise when it reduces the number of characteristics.

5. Discussion

In general, the three methods discussed in this paper can provide precise estimations. The first exercise shows the random forest (200) method to be more accurate, followed by the conventional hedonic model. The first two methods do not allow us to identify how or why some housing-related characteristics influence price increases but they do provide precise values in the context of the data. In the third exercise, the quantile hedonic model provides less precision in predicting housing prices (the errors are larger) but it explains why prices evolve and, with a small, precise redesign of the model, it would capable of predicting housing prices for the sample and also provide more accurate forecasts.

Debates over the lack of precision in the hedonic method point to changes in expectations or tastes, which are led by other variables. Such relationships can be associated with complex non-linearities the hedonic method has difficulty identifying and which have not yet been solved. Nevertheless, the machine learning method can support the identification of non-linearities. As a secondary result of the random forest that was adapted for the purpose of this study, a three-dimensional relationship was defined among the main housing characteristics that explain housing prices. The RF method allows us to infer how the

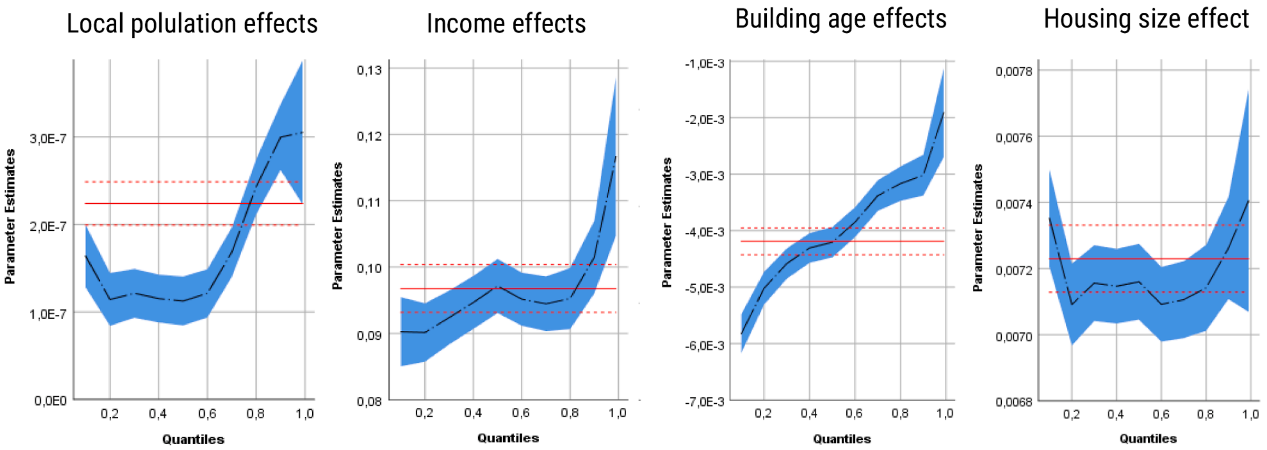


Fig. 7. Plots of the effects on the different variables. The blue shadow measures the confidence intervals of the estimated parameters (on the black line). The red line gives the parameter estimates of the ordinary linear squares regression and the dotted red lines indicate their interval bounds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

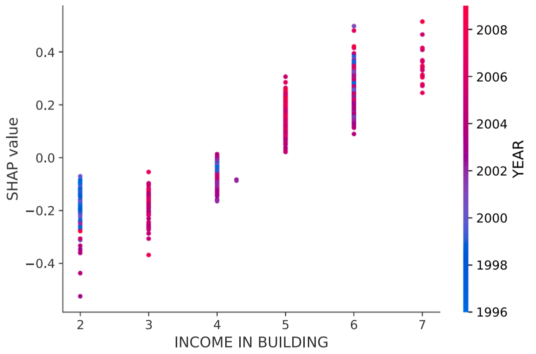


Fig. 9. Dependence plot with variable INCOME IN BUILDING against impact value using YEAR colour scale. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

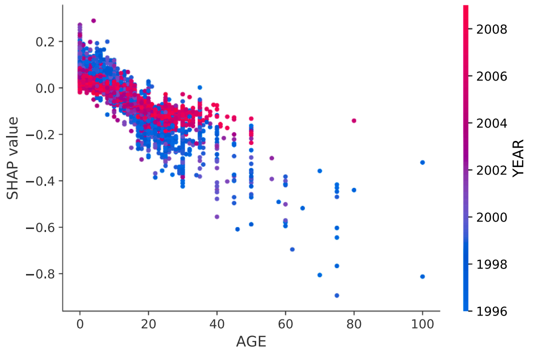


Fig. 11. Dependence plot with the variable AGE against the impact value using the YEAR colour scale. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

relationship between two variables evolves to being capable of explaining housing prices. The follow figures[10] shows four examples: HOUSING SIZE and LOCATION (region in the province where the house is located) (Fig. 8), INCOME BUILDING and YEAR (Fig. 9), CONSTRUCTION QUALITY and AGE (Fig. 10); AGE and YEAR (Fig. 11). These are the non-linearities that are traditionally debated in the literature.

Regarding the first relationship in Fig. 8, the ‘x’ value is the housing size (m2) and the ‘y’ is the location. The order of the location is from the North to the South and from the coast to the interior. The Shapley value identifies how the influence on prices rises as with the house size, but in a non-linear way (as the literature has identified). The blue and red points are together since it is common for the first locations (north of Alicante) to have 100 m2 in area in which to assert their influence. However, if the house is located in the Southern parts of the province, then as the house sizes increase, the influence larger dwellings have on housing prices decreases. In the figure, this can be seen in the way the blue points differ from the red points at 250–300 m2. The non-linearity captures the different effects an extra m2 has on housing prices if these dwellings are built in different regions.

The second non-linearity is related to the influence of income and year on housing prices. The Fig. 9 explains how income affects housing prices in a range of years. Income presents both positive and negative effects on housing prices. The effect is negative when income is low (up to level 4) and positive when income is high (from level 5 up). Interestingly, the random forest identifies the negative effect of low income on housing prices for 2004, onward. But it only identifies the effects of very low income during the period 1996–2000. These findings support the idea of that low-income households reduce housing prices while higher-income households increase these prices, thus, defining a segmentation of the housing-price dynamics by level of income.

The third non-linearity relates to the construction quality and income level (Fig. 10). The accumulation of red points on the left-hand side of the figure is associated with lower housing prices. This supports the accepted idea that older buildings are associated with lower construction quality. The blue points in the level-3 construction quality suggest that newer buildings also present low quality and, thus, negatively affect prices. In this sense, this non-linearity agrees with what is commonly thought in the economics literature.

The last non-linearity (Fig. 11) explains how building age affects housing prices, depending on the year (cycle). The red colour identifies how newer buildings had null effect on prices from 2004 and onward whereas they had positive affects during the previous periods. After 20 years, the effect on housing prices was mostly negative during the first part of the observed period (1996–2000) but this effect was stabilised over the next few years. These figures suggest that the impact of building age differs depending on the time the house is supplied on the market, clearly capturing the changes in owner expectations.

It is possible to see these results by using the random forest method;

[10] The interpretations of these figures are not straight forward. The Shapley value measures the influence of the combined variables on prices; the red colour identifies the effects on the variables on the ‘x’ axis when the ‘y’ variables show larger values (in red).

Table 4
Ranking of main explanatory features related to housing prices.

Ranking	Random forest (200)	Hedonic regression	Hedonic through origin
1	Year	Housing size m2	Postal code
2	Housing size m2	Location	Age
3	Income building	Year	Urban type
4	Age	Income building	Population
5	Number of bathrooms	Income level	Economic activity
6	Location	Number of bathrooms	Population growth
7	Housing type t	Age of neighbourhood	Urban/rural
8	Number lifts	Economic activity	Type of housing use
9	Views	Road quality	Income level
10	Economic activity	Number of lifts	Population density
11	Construction q	Religious facility	Population develop
12	Population	Views	Road quality
13	Urbanization q	Age	Sewer type
14	Number of dwellings in building	Type of housing use	Sewer quality
15	Outdoor living space	Month	Lighting system type
16	Retail facilities q neighbour	Housing use	Lighting system quality
17	Orientation	Number of dwellings in building	retail facilities q
18	Population growth	Population	School facilities q
19	Housing use	Exterior rooms	Religious facility
20	Postal code	Health facilities q	Sports facilities q
21	Provincial code	Outdoor living space	Health facilities q
22	Income level	Population develop	Bus
23	Population density	Sports facilities q	Train
24	Age of neighbourhood	Retail facilities q neighbour	Underground
25	Longitude	Housing type t	Housing type t
26	Month	Construction q	Number of dwellings in building
27	Population development	Urban type	Number of lifts
28	Population density	Train	Age of neighbourhood
29	Number of rooms	School facilities q	Retail facilities q neighbour
30	Retail facilities q	Urbanization q	Income building
31	Exterior rooms	Longitude	Population density neighbour
32	Type housing use	Lighting system quality	Location

as for hedonics, this requires much more complexity. Nevertheless, showing the results visually could serve as an economic interpretation if there were a way for hedonic models to clearly capture the effects in three dimensions.

The whole empirical exercise shows how the Random forest method is capable of capturing and visualizing changes in data patterns (see figure 8 or 11) hidden in data and key for producing accurate predictions. Non-linearities identification has been a way followed by economic researchers to capture how the market mechanism is performing in a particular time or space, so any knowledge that could develop an algorithm allowing to precisely model non-linearities into causal models would be a significant step forward in socio-economic analysis.

For the time being, ML is not capable of explaining why the non-linearities happen. In fact, ML methods cannot clarify why the identified patterns change and the effect of any omitted variable or hidden factors in the dataset, for instance. Although RF can categorize the variables mainly affecting the prediction, it is still not capable of explaining why and how. The accuracy almost depends on the data

Table 5
Results marked with asterisks mean that *** p-value < 0.01 and ** p-value < 0.05.

Parameter (quantile)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Constant	10.3***	10.46***	10.47***	10.507***	10.509***	10.502***	10.472***	10.367***	10.235***
Postal code	−0.00012***	−0.000120***	−0.00010***	−0.000089***	−0.000092***	−0.000081***	−0.000056***	−0.000014***	−0.000086***
Age	−0.0026***	−0.0022***	−0.0020***	−0.0018***	−0.0016***	−0.0016***	−0.0015***	−0.0013***	−0.0011***
Urban type	−0.0051	−0.0044	−0.0097***	−0.0114***	−0.0118***	−0.0139***	−0.0156***	−0.0193***	−0.00242***
Population	1.638E−07***	1.143E−07***	1.213E−07***	1.153E−07***	1.125E−07***	1.213E−07***	1.684E−07***	2.432E−07***	2.998E−07***
Economic activity	0.016***	0.017***	0.019***	0.021***	0.021***	0.021***	0.020***	0.019***	0.017***
Population growth	−0.022***	−0.019***	−0.020***	−0.019***	−0.021***	−0.023***	−0.031***	−0.040***	−0.049***
Urban rural	0.0628**	0.0614***	0.0623***	0.0684***	0.0544***	0.0486**	0.0401**	0.04	0.0228
Type housing use	0.0277***	0.0340***	0.0357***	0.0399***	0.0436***	0.0447***	0.0487***	0.0534***	0.0622***
Income level	0.0765***	0.0802***	0.0828***	0.0838***	0.0849***	0.0854***	0.0869***	0.0902***	0.0996***
Population density	0.0094**	0.0115***	0.0102***	0.0073**	0.0048	0.0012	−0.0011	−0.0082**	−0.0064
POPULATION DEVELOP	−0.0328***	−0.0277***	−0.0239***	−0.0224***	−0.0218***	−0.0226***	−0.0212**	−0.0196***	−0.0209***
Road quality	0.0512	0.0073	0.0295	0.0792***	0.0771***	0.0677***	0.0973***	0.0743**	0.0974***
Sewer type	−0.0263**	−0.0253**	−0.0299***	−0.0280***	−0.0124	−0.0028	0.0017	0.00	0.0183
Lighting system type	0.0392***	0.0201	0.0254**	0.0149	0.0092	−0.0028	−0.0044	0.01	0.0034
Lighting system quality	−0.0066	0.0353	0.0112	−0.0362	−0.0315	−0.0229	−0.0575**	−0.04	−0.0689**
Retail facilityes Q	0.0079***	0.0049***	0.0043***	0.0064***	0.0078***	0.0077***	0.0092***	0.0099***	0.0138***
School facilities Q	−0.0125	−0.0279***	−0.0328***	−0.0343***	−0.0367***	−0.0328***	−0.0363***	−0.0308***	−0.0307***
Religious facility	−0.0583***	−0.0484***	−0.0520***	−0.0548***	−0.0571***	−0.0594***	−0.0649***	−0.0712***	−0.0607***
Sports facilities Q	0.0234***	0.0173***	0.0192***	0.0224***	0.0214***	0.0283***	0.0353***	0.0429***	0.0374***
Health facilities Q	0.0198**	0.0256***	0.0324***	0.0324***	0.0388***	0.0321***	0.0359***	0.0296***	0.0367***
Bus	−0.0107***	−0.0079***	−0.0081***	−0.0077***	−0.0068***	−0.0040**	−0.0049***	−0.0051***	−0.0083***
Train	0.0170***	0.0205***	0.0225***	0.0214***	0.0213***	0.0223***	0.0222***	0.0203***	0.0180***
Underground	0.0073	−0.0021	0.0086	0.0024	−0.0059	−0.0156	−0.0070	−0.01	0.0
Housing type T	0.0152***	0.0170***	0.0183***	0.0163***	0.0121***	0.0113***	0.0087***	0.00	0.0

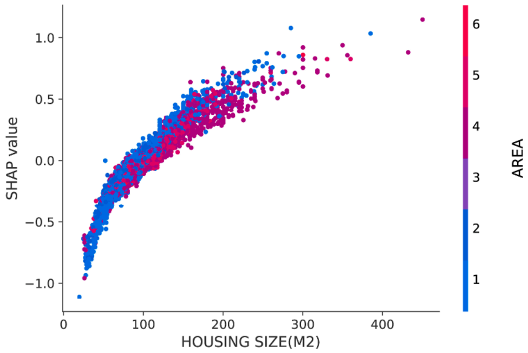


Fig. 8. Dependence plot with the variable HOUSING SIZE(M2) against the impact value using a colour scale for LOCATION.

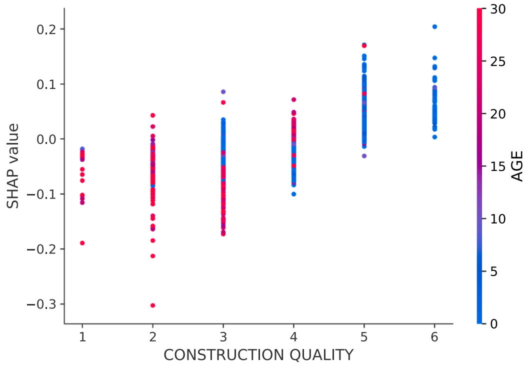


Fig. 10. Dependence plot with variable CONSTRUCTION QUALITY against the impact value using a color scale for LOCATION. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

managed and how quick the new information is added for the analysis. These raise issues related to computation costs and massive datasets to be managed as well as the prior beliefs/theories to be accounted for as an alternative ingredient of ML method application.

Another area where machine learning techniques are not developed enough is in the capacity to predict the future, as the current methods predict the present. From a socio-economic point of view, the ML techniques could give good predictions to satisfy the information requirements of particular markets with fast and multiple transactions, like the financial markets, but not in the case of others more longer-term based. The housing market belongs to the latter, and the interest to advance their future price evolution is considerable because of their effects on several socio-economic dimensions like social life or the macro-economic stability, to mention only two examples. Saying that, ML methods are used in the massive valuation tools in the real estate market, but the prediction is still a remaining issue. The current treatment of time in ML as an additional variable in the dataset ignores the time dimension of the extracted patterns related to the other variables. As in the non-linearities case, ML can merge with econometric-time series knowledge to develop tools which can make more accurate predictions for next periods. Predicting the future of data is one of the remaining challenges for artificial intelligence.

6. Conclusion and recommendations for future study

This research presents the results of several predictive models that are based on two different methodologies: machine learning and hedonic regressions. A total of seven different models were tested for their predictive capabilities, using a large micro database of housing (asking) prices in Alicante City, Spain, for the period 1996–2012. The first group of methods tested a range of machine learning (ML) tools to calculate the most relevant components that explain housing prices. ML tools provide high predictive capability and show how the random forest (RF) method provides superior predictions with higher explanatory power due to the manner in which it classifies the components that are used to estimate housing prices. However, these methods are somewhat weak in terms of explaining the rationale behind price levels. However, the more advanced RF method is capable of identify those characteristics that contribute to the main part of a house's value. The findings from using RF agree with the literature where it notes that location and time are the crucial features for determining prices. A secondary result from the RF method is that it allows us to identify and visualise the discontinuities in the relationships between the housing characteristics, whereas these relationships should normally be non-linear. That is, these relationship discontinuities may affect a particular characteristic and the impact it might have on housing prices, depending on whether a third characteristic also has affects on the housing price. This is a clear non-linear relationship that reflects a hidden causality which can be observed in three dimensions by using the RF method but is difficult to capture using other methods. The hedonic regression method is estimated using two tools: The first is the ordinary least squares hedonic method, which starts from the hypothesis that the effects of each characteristic (estimated parameter) is constant. The second tool is the quantile hedonic regression method which considers that each characteristic could present a non-linear effect over prices and estimate such an effect by each 10% quantile. The results of both methods are fully consistent with the literature whose main results rank the more relevant variables that determine housing prices. The quantile results identify the non-linearities in the way certain characteristics affect housing prices, depending on their ranking. Both results are compared and Table A-1 identifies the relevance of the characteristics as they are measured using the two above techniques. In terms of accuracy, the ML methods are superior to all of the regression methods, although their capacity to explain a socioeconomic variable can be found to be spurious when the variables are highly correlated. In terms of the effects on quantification, the regression models are superior as they are capable of precisely identifying each characteristic's particular effect and defining the range of effects. We find that the RF method is superior in terms of identifying the non-linearities. In fact, this method can visualize the changing effects of three different attributes that show the non-linearities that are potentially responsible of an asymmetric effect of a particular characteristic on the price determination. However, the RF does not quantify such non-linear relationships. The quantile regression is capable of capturing two-dimensional non-linearities; that is, it can capture the asymmetry in the housing-price reactions of a particular attribute, depending on the price (quantile) levels; it can also provide a precise quantification of each of those impacts that are associated to the price quantile. The latter explains the observed asymmetric housing-price reactions; but, at the moment, it is not possible to capture the three-dimensional effects which can be hidden within non-observable causal links among the housing attributes.

This paper is the first to provide a comparison of the two methodologies using real data and it opens this field up for further research. The development of emerging tools that can identify the three dimensions that have causal hidden relationships on housing prices is crucial to understanding how these prices react in each market.

As a proposal for future work could be raised to apply techniques based on deep artificial neural networks (Deep Learning) that are currently being applied in complex problems getting very good results. For this new approach would be a challenge in terms of explainability of the model, since these complex neural networks have more limited explanatory techniques than those used in the current article.

CRediT authorship contribution statement

Juan Ramón Rico-Juan: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **Paloma Taltavull de La Paz:** Conceptualization, Methodology, Software, Data curation, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Machine learning numerical results

The tables below show the numerical results for each metric obtained from the machine learning algorithms. The algorithms are sorted by name. The column that shows the average percentages also presents in parentheses, in the cases of the RMSE and MAE metrics, the relationship between the average value and the average value of the PRICE variable defined in Eqs. 4 and 6, respectively. In addition to the averages obtained, the results of each experiment are also shown within the 10-fold cross validation technique.

A.1. Root mean squared error

.

Algorithm	Average (%)	10-CV
AdaBoost	0.354 (3.1)	[0.36, 0.35, 0.36, 0.35, 0.36, 0.35, 0.35, 0.35, 0.36, 0.35]
CatBoost	0.161 (1.4)	[0.16, 0.16, 0.16, 0.16, 0.16, 0.16, 0.16, 0.16, 0.17, 0.16]
Decision Tree	0.310 (2.7)	[0.31, 0.31, 0.31, 0.31, 0.31, 0.31, 0.31, 0.31, 0.31, 0.31]
Linear Lasso	0.613 (5.3)	[0.62, 0.61, 0.61, 0.61, 0.62, 0.61, 0.62, 0.61, 0.61, 0.61]
Linear Regression	0.259 (2.2)	[0.26, 0.26, 0.26, 0.25, 0.26, 0.26, 0.26, 0.26, 0.26, 0.26]
Linear Ridge	0.259 (2.2)	[0.26, 0.26, 0.26, 0.25, 0.26, 0.26, 0.26, 0.26, 0.26, 0.26]
MLP	0.206 (1.8)	[0.20, 0.21, 0.20, 0.20, 0.21, 0.21, 0.22, 0.20, 0.21, 0.20]
Nearest Neighbours (1)	0.172 (1.5)	[0.17, 0.17, 0.17, 0.17, 0.17, 0.17, 0.18, 0.17, 0.18, 0.17]
Nearest Neighbours (3)	0.163 (1.4)	[0.16, 0.17, 0.16, 0.16, 0.16, 0.16, 0.17, 0.16, 0.17, 0.16]
Nearest Neighbours (5)	0.168 (1.5)	[0.17, 0.17, 0.16, 0.17, 0.17, 0.16, 0.17, 0.16, 0.18, 0.17]
Nearest Neighbours (7)	0.173 (1.5)	[0.17, 0.18, 0.17, 0.17, 0.17, 0.17, 0.18, 0.17, 0.18, 0.17]
Random Forest (100)	0.086 (0.7)	[0.08, 0.10, 0.08, 0.08, 0.08, 0.09, 0.09, 0.08, 0.10, 0.08]
Random Forest (200)	0.078 (0.7)	[0.07, 0.09, 0.07, 0.07, 0.07, 0.08, 0.09, 0.08, 0.09, 0.07]
Random Forest (300)	0.077 (0.7)	[0.07, 0.09, 0.07, 0.07, 0.07, 0.08, 0.09, 0.07, 0.09, 0.07]
Random Forest (400)	0.077 (0.7)	[0.07, 0.09, 0.07, 0.07, 0.07, 0.08, 0.09, 0.07, 0.09, 0.07]
Random Forest (500)	0.077 (0.7)	[0.07, 0.09, 0.07, 0.07, 0.07, 0.08, 0.09, 0.07, 0.09, 0.07]
XGBRegressor	0.200 (1.7)	[0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20]

A.2. Mean absolute error

.

Algorithm	Average (%)	10-CV
AdaBoost	0.272 (2.4)	[0.28, 0.27, 0.28, 0.27, 0.27, 0.27, 0.27, 0.27, 0.27, 0.27]
CatBoost	0.120 (1.0)	[0.12, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12, 0.12]
Decision Tree	0.234 (2.0)	[0.23, 0.24, 0.24, 0.23, 0.23, 0.24, 0.24, 0.23, 0.23, 0.23]
Linear Lasso	0.490 (4.3)	[0.49, 0.49, 0.49, 0.49, 0.49, 0.49, 0.49, 0.49, 0.49, 0.49]
Linear Regression	0.200 (1.7)	[0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20]
Linear Ridge	0.200 (1.7)	[0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20]
MLP	0.155 (1.3)	[0.16, 0.15, 0.16, 0.15, 0.16, 0.16, 0.16, 0.15, 0.15, 0.15]
Nearest Neighbours (1)	0.076 (0.7)	[0.08, 0.07, 0.07, 0.08, 0.07, 0.08, 0.08, 0.07, 0.08, 0.08]
Nearest Neighbours (3)	0.090 (0.8)	[0.09, 0.09, 0.09, 0.09, 0.09, 0.09, 0.09, 0.09, 0.09, 0.09]
Nearest Neighbours (5)	0.100 (0.9)	[0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10]
Nearest Neighbours (7)	0.110 (1.0)	[0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11]
Random Forest (100)	0.040 (0.3)	[0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04]
Random Forest (200)	0.030 (0.3)	[0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03]
Random Forest (300)	0.030 (0.3)	[0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03]
Random Forest (400)	0.030 (0.3)	[0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03]
Random Forest (500)	0.030 (0.3)	[0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03]
XGBRegressor	0.150 (1.3)	[0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15]

A.3. R square score

.

Algorithm	Average	10-CV
AdaBoost	0.667	[0.66, 0.66, 0.66, 0.68, 0.67, 0.67, 0.68, 0.67, 0.66, 0.66]
CatBoost	0.935	[0.94, 0.93, 0.94, 0.94, 0.94, 0.93, 0.93, 0.93, 0.93, 0.94]
Decision Tree	0.747	[0.75, 0.74, 0.75, 0.75, 0.75, 0.75, 0.74, 0.75, 0.74, 0.75]
Linear Lasso	0.423	[0.42, 0.42, 0.43, 0.43, 0.43, 0.42, 0.42, 0.42, 0.42, 0.42]
Linear Regression	0.823	[0.82, 0.82, 0.83, 0.83, 0.83, 0.82, 0.82, 0.82, 0.82, 0.82]
Linear Ridge	0.823	[0.82, 0.82, 0.83, 0.83, 0.83, 0.82, 0.82, 0.82, 0.82, 0.82]
MLP	0.886	[0.89, 0.88, 0.89, 0.89, 0.89, 0.89, 0.87, 0.89, 0.88, 0.89]
Nearest Neighbours (1)	0.923	[0.93, 0.92, 0.93, 0.92, 0.93, 0.92, 0.92, 0.92, 0.92, 0.92]
Nearest Neighbours (3)	0.928	[0.93, 0.93, 0.93, 0.93, 0.93, 0.93, 0.92, 0.93, 0.92, 0.93]
Nearest Neighbours (5)	0.927	[0.93, 0.92, 0.93, 0.93, 0.93, 0.93, 0.92, 0.93, 0.92, 0.93]
Nearest Neighbours (7)	0.919	[0.92, 0.92, 0.92, 0.92, 0.92, 0.92, 0.92, 0.92, 0.91, 0.92]
Random Forest (100)	0.978	[0.98, 0.97, 0.98, 0.98, 0.98, 0.98, 0.98, 0.98, 0.97, 0.98]
Random Forest (200)	0.985	[0.99, 0.98, 0.99, 0.99, 0.99, 0.98, 0.98, 0.98, 0.98, 0.99]
Random Forest (300)	0.986	[0.99, 0.98, 0.99, 0.99, 0.99, 0.98, 0.98, 0.99, 0.98, 0.99]
Random Forest (400)	0.986	[0.99, 0.98, 0.99, 0.99, 0.99, 0.98, 0.98, 0.99, 0.98, 0.99]
Random Forest (500)	0.986	[0.99, 0.98, 0.99, 0.99, 0.99, 0.98, 0.98, 0.99, 0.98, 0.99]
XGBRegressor	0.897	[0.90, 0.89, 0.90, 0.90, 0.90, 0.90, 0.89, 0.90, 0.89, 0.90]

A.4. Precision

.

Algorithm	Average	10-CV
AdaBoost	0.728	[0.72, 0.73, 0.72, 0.73, 0.73, 0.73, 0.73, 0.73, 0.73, 0.73]
CatBoost	0.880	[0.88, 0.88, 0.88, 0.88, 0.88, 0.88, 0.88, 0.88, 0.88, 0.88]
Decision Tree	0.766	[0.77, 0.76, 0.76, 0.77, 0.77, 0.76, 0.76, 0.77, 0.77, 0.77]
Linear Lasso	0.510	[0.51, 0.51, 0.51, 0.51, 0.51, 0.51, 0.51, 0.51, 0.51, 0.51]
Linear Regression	0.800	[0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80]
Linear Ridge	0.800	[0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80, 0.80]
MLP	0.845	[0.84, 0.85, 0.84, 0.85, 0.84, 0.84, 0.84, 0.85, 0.85, 0.85]
Nearest Neighbours (1)	0.924	[0.92, 0.93, 0.93, 0.92, 0.93, 0.92, 0.92, 0.93, 0.92, 0.92]
Nearest Neighbours (3)	0.910	[0.91, 0.91, 0.91, 0.91, 0.91, 0.91, 0.91, 0.91, 0.91, 0.91]
Nearest Neighbours (5)	0.900	[0.90, 0.90, 0.90, 0.90, 0.90, 0.90, 0.90, 0.90, 0.90, 0.90]
Nearest Neighbours (7)	0.890	[0.89, 0.89, 0.89, 0.89, 0.89, 0.89, 0.89, 0.89, 0.89, 0.89]
Random Forest(100)	0.960	[0.96, 0.96, 0.96, 0.96, 0.96, 0.96, 0.96, 0.96, 0.96, 0.96]
Random Forest (200)	0.970	[0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97]
Random Forest (300)	0.970	[0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97]
Random Forest (400)	0.970	[0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97]
Random Forest (500)	0.970	[0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97]
XGBRegressor	0.850	[0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85]

Appendix B. Hedonic estimation model

Hedonic model estimation. Differences in parameters by quantile (Quantile Regressors) (Table 4.

References

Abreu, D., & Brunnermeier, M. K. (2002). Synchronization risk and delayed arbitrage. *Journal of Financial Economics, 66*, 341–360.

Adair, A. S., Berry, J. N., & McGreal, W. S. (1996). Hedonic modelling, housing submarkets and residential valuation. *Journal of Property Research, 13*, 67–83.

Allen, F., & Gale, D. (1998). Optimal financial crises. *The Journal of Finance, 53*, 1245–1284.

Anselin, L. (2013). *Spatial econometrics: Methods and models* (Vol. 4). Springer Science & Business Media.

Black, A., Fraser, P., & Hoesli, M. (2006). House prices, fundamentals and bubbles. *Journal of Business Finance & Accounting, 33*, 1535–1555.

Bourassa, S. C., Hoesli, M., Scognamiglio, D., & Zhang, S. (2011). Land leverage and house prices. *Regional Science and Urban Economics, 41*, 134–144.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Case, B., Colwell, P. F., Leishman, C., & Watkins, C. (2006). The impact of environmental contamination on condo prices: a hybrid repeat-sale/hedonic approach. *Real Estate Economics, 34*, 77–107.

Case, B., Wachter, S. & et al. (2005). Residential real estate price indices as financial soundness indicators: Methodological issues. BIS paper (pp. 197–211).

Case, K. E. & Shiller, R. J. (2003). Is there a bubble in the housing market? Brookings papers on economic activity, 2003, 299–362.

Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. CoRR, abs/1603.02754.

Clapham, E., Englund, P., Quigley, J. M., Redfearn, C. L. & et al. (2004). Revisiting the past: revision in repeat sales and hedonic indexes of house prices. Lusk Center for Real Estate Working Paper.

Clapp, J., & Giaccotto, C. (2002). Evaluating house price forecasts. *Journal of Real Estate Research, 24*, 1–26.

Clapp, J. M. (2003). A semiparametric method for valuing residential locations: application to automated valuation. *The Journal of Real Estate Finance and Economics, 27*, 303–320.

Coulson, N. (2012). House price index methodologies. In *International encyclopedia of housing and home*.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*, 21–27.

Dorogush, A.V., Ershov, V. & Gulin, A. (2018). Catboost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

Dougherty, A., & Van Order, R. (1982). Inflation, housing costs, and the consumer price index. *The American Economic Review, 72*, 154–164.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*, 119–139.

Geltner, D. (1993). Estimating market values from appraised values without assuming an efficient market. *Journal of Real Estate Research, 8*, 325–345.

Goh, K. C., Seow, T. W., & Goh, H. H. (2013). Challenges of implementing sustainability in malaysian housing industry. In *International conference on sustainable built environment for now and the future (SBE2013)*.

Goodman, A. C., & Thibodeau, T. G. (1995). Age-related heteroskedasticity in hedonic house price equations. *Journal of Housing Research*, 25–42.

Goodman, A. C., & Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics, 12*, 181–201.

Haurin, D. R. (1991). Income variability, homeownership, and housing demand. *Journal of Housing Economics, 1*, 60–74.

Hinton, G. E. (1990). Connectionist learning procedures. In *Machine learning* (Vol. III, pp. 555–610). Elsevier.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55–67.

Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy, 82*, 657–673.

Hwang, M., & Quigley, J. M. (2004). Selectivity, quality adjustment and mean reversion in the measurement of house values. *The Journal of Real Estate Finance and Economics, 28*, 161–178.

Hwang, M., & Quigley, J. M. (2006). Economic fundamentals in local housing markets: Evidence from us metropolitan regions. *Journal of Regional Science, 46*, 425–453.

Kauko, T. (2003). On current neural network applications involving spatial modelling of property prices. *Journal of Housing and the Built Environment, 18*, 159–181.

Kauko, T. & d'Amato, M. (2009). Book review mass appraisal methods: An international perspective for property valuers.

Kim, K.-H., & Suh, S. H. (1993). Speculation and price bubbles in the korean and japanese real estate markets. *The Journal of Real Estate Finance and Economics, 6*, 73–87.

Taltavull de La Paz, P., López, E., & Juárez, F. (2017). Ripple effect on housing prices. Evidence from tourist markets in alicante, spain. *International Journal of Strategic Property Management, 21*, 1–14.

Linneman, P. (1980). Some empirical results on the nature of the hedonic price function for the urban housing market. *Journal of Urban Economics, 8*, 47–68.

Lundberg, S. (2019). Shap (shapley additive explanations). https://github.com/slundberg/shap.

Lundberg, S. M., & Lee, S.-I. (2017). *Consistent feature attribution for tree ensembles*. arXiv preprint arXiv:1706.06060.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765–4774).

McGreal, S., & Taltavull de La Paz, P. (2012). An analysis of factors influencing accuracy in the valuation of residential properties in spain. *Journal of Property Research, 29*, 1–24.

Meen, G. (1999). Regional house prices and the ripple effect: A new interpretation. *Housing Studies, 14*, 733–753.

Meen, G. (2012). *Modelling spatial housing markets: Theory, analysis and policy* (Vol. 2). Springer Science & Business Media.

Meen, G., & Andrew, M. (2004). On the use of policy to reduce housing market segmentation. *Regional Science and Urban Economics, 34*, 727–751.

Mikhed, V., & Zemčík, P. (2009). Do house prices reflect fundamentals? Aggregate and panel data evidence. *Journal of Housing Economics, 18*, 140–149.

Muellbauer, J., & Murphy, A. (1997). Booms and busts in the uk housing market. *The Economic Journal, 107*, 1701–1727.

Ortalo-Magne, F., & Rady, S. (2006). Housing market dynamics: On the contribution of income shocks and credit constraints. *The Review of Economic Studies, 73*, 459–485.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert Systems with Applications, 42*, 2928–2934.

Peek, J., & Wilcox, J. A. (1991). The baby boom, pent-up demand, and future house prices. *Journal of Housing Economics, 1*, 347–367.

Poterba, J. M. (1984). Tax subsidies to owner-occupied housing: An asset-market approach. *The Quarterly Journal of Economics, 99*, 729–752.

Poterba, J. M., Weil, D. N., & Shiller, R. (1991). House price dynamics: The role of tax policy and demography. *Brookings Ppapers on Economic Activity, 1991*, 143–203.

Riddel, M. (1999). Fundamentals, feedback trading, and housing market speculation: Evidence from california. *Journal of Housing Economics, 8*, 272–284.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*, 34–55.

Roth, A. E. (1988). *The Shapley value: Essays in honor of Lloyd S*. Shapley: Cambridge University Press.

Saiz, A. (2010). The geographic determinants of housing supply. *The Quarterly Journal of Economics, 125*, 1253–1296.

Stiglitz, J. E. (1993). The role of the state in financial markets. *The World Bank Economic Review, 7*, 19–52.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*, 267–288.

Weisberg, S. (2005). *Applied linear regression* (Vol. 528). John Wiley & Sons.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*, 80–83.

Yoo, S., Im, J., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in onondaga county, ny. *Landscape and Urban Planning, 107*, 293–306.

Zhao, Q., & Hastie, T. (2019). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 1–10.