Decision Analytics Journal 6 (2023) 100166

Contents lists available at ScienceDirect

Decision Analytics Journal

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

A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization

Salim Lahmiri [a,*], Stelios Bekiros [b,c], Christos Avdoulas [d,e]

[a] Department of Supply Chain & Business Technology Management, John Molson School of Business, Concordia University, Montreal, Canada
[b] FEMA, University of Malta, MSD 2080 Msida, Malta
[c] LSE Health, London School of Economics and Political Science (LSE), London WC2A2AE, UK
[d] Athens University of Economics and Business, Athens, Greece
[e] Inter-American Research Center, ACHMEA, Canada

ARTICLE INFO

Keywords:
House price prediction
Predictive analytics
Boosting ensemble regression trees
Support vector regression
Gaussian process regression
Bayesian optimization

ABSTRACT

The valuation of house prices is drawing noteworthy attention due to worldwide financial and real estate crises in the last decade. Therefore, there is an immediate need to design more effective predictive systems of house prices. Indeed, investors, creditors, and governments are all interested in such predictive systems to improve their buying and lending decisions and activities. This study explores the application of artificial intelligence, machine learning, and nonlinear statistical models to house price prediction problems. In that order, we use boosting ensemble regression trees, support vector regression, and Gaussian process regression. Bayesian optimization is implemented in a ten-fold cross-validation framework to determine their respective optimal kernels and parameter values. Four performance metrics are used to evaluate the prediction ability of each predictive system. The experimental results showed that boosting ensemble regression trees performed the best, followed by Gaussian process regression and support vector regression. In addition, all three aforementioned predictive systems outperformed artificial neural networks and multi-variate regression employed in recent work on the same data set. Under this perspective, it is concluded that boosting ensemble regression trees are clear candidates to be considered for operational house price prediction in Taiwan.

1. Introduction

House price evaluation is an important topic in the financially attractive domain of real estate valuation and management since it can help real estate companies, financial institutions, and investors negotiate price and take appropriate actions in advance. Recent subprime mortgage crisis and great recession which have occurred across world financial markets during the late 2000s and early 2010s caused economic decline worldwide, specifically in the USA economy. Indeed, financial and economic globalization generated waves of economic distress across societies and national economies; thus, many Americans endured negative financial impact. Since then, appropriate house price evaluation has become a public concern and expert real estate advice is definitively needed for homebuyers, sellers, financial institutions, and government.

Surely, appropriate house pricing is a very important topic in both practical and academic fields of real estate finance. From practical view, homebuyers, sellers, creditors, senior management, and auditors are all interested in house price evaluation for the reason that it has great impact on their financial and investment decision making.

Indeed, to make, the right decision on whether to buy or sell a house, an economic agent (including for instance a homebuyer, seller, creditor, senior management, and auditor) need to use an appropriate predictive model to predict the accurate value of house price. In other words, he needs an accurate model for house evaluation. Indeed, it is important to value a property for a purchase to be able to generate profit. For instance, an accurate model is need by homebuyer to evaluate his investment, by seller to evaluate profit, by creditor to evaluate risk, and by senior manager and auditor to better manage assets portfolio.

More to the point, real estate and mortgage crisis also brings serious social problems such as unemployment, economic depression and financial crisis if many companies run into financial distress in the same period. Consequently, there is insistent demand for accurate house price evaluation technical models in practice, to which many scholars have been devoted.

Indeed, driven by the strong business needs, many statistical models have been proposed for house price evaluation in the past few years.

* Corresponding author.
E-mail addresses: salim.lahmiri@concordia.ca (S. Lahmiri), stelios.bekiros@um.edu.mt, s.bekiros@lse.ac.uk (S. Bekiros), chr.abdoulas@gmail.com (C. Avdoulas).

https://doi.org/10.1016/j.dajour.2023.100166
Received 22 November 2022; Received in revised form 14 January 2023; Accepted 15 January 2023
Available online 18 January 2023

For instance, various statistical models have been employed; including regression analysis [1], semi-parametric regression [2], large-scale Bayesian vector autoregressive model [3], Granger causality and variance decomposition [4], lognormal regression model [5], smooth transition model and error correction models [6], analogical regression [7], and dynamic model averaging and dynamic model selection [8,9].

In recent years, artificial intelligence and machine learning based systems and algorithms are attracting more attention than conventional statistical models in house price evaluation. Indeed, this growing interest is due to the fast development of computer power and data storage technologies and their respective ability to provide high prediction accuracies; thus, increasing profits and decreasing losses. In this regard, artificial intelligence systems and machine learning models include fuzzy logic system [10], hybrid fuzzy regression-fuzzy cognitive map algorithm [11], adaptive neuro-fuzzy system [12], support vector machine optimized by particle swarm optimization [13], repeated incremental pruning to produce error reduction (RIPPER) algorithm [14], combination of ensemble empirical mode decomposition and support vector regression [15], and case-based reasoning [16].

For instance, a fuzzy logic system was employed to predict house selling price in different regions of Eskişehir city in Turkey by using house, environmental, transportation, and regional socio-economic factors [10]. It was concluded that the predictions are very close to the unit real price values. A hybrid algorithm based on fuzzy linear regression and fuzzy cognitive map was proposed to deal with imprecise and ambiguous inputs (for example, various supply and demand factors) to better forecast house price in Iran [11]. It was concluded that the proposed hybrid system is effective in presence of uncertainty and severe noise associated with the housing market. Gerek [12] compared ANFIS with grid partition (GP) and ANFIS with sub clustering (SC) in predicting house price in the construction sector in southern Turkey by using exclusively industry factors. The simulations results showed that ANFIS-GP system was, to a small degree, better than the ANFIS-SC system. In [13], the authors used the support vector machine to predict house average selling price in China by using previous average selling price as inputs. The parameters of the SVM was tuned by either grid algorithm, genetic algorithm or particle swarm optimization. They found that the SVM tuned by particle swarm optimization outperformed backpropagation neural networks, SVM tuned by grid algorithm and SVM tuned by genetic algorithm. In [14], the authors employed repeated incremental pruning to produce error reduction (RIPPER) algorithm trained with 28 variables selected by stepwise logistic regression to predict housing price in the United States. The RIPPER algorithm outperformed C4.5 algorithm, Naïve Bayes, and AdaBoost algorithm. Besides, the authors in [15] combined ensemble empirical mode decomposition (EEMD) and support vector regression to predict sudden house price drops in the United States. The presented model was trained with ten annual macroeconomic variables. The experimental results showed that the presented approach outperformed random walk, Bayesian autoregressive, and Bayesian vector autoregressive model. In [16], the authors found that artificial neural networks outperform the multivariate regression model in forecasting house price in Taiwan.

Other studies focused on decision trees for model and predict house price. For instance, in [17], random forest algorithm was employed to predict House Price Index in United States and achieved a ±5% error margin. In addition, decision trees, gradient-boosting and random forest algorithm were found to be effective compared to multiple linear regression model when applied to Australian market data [18]. Finally, artificial neural networks were found to be effective in predicting house price in China (China Real Estate Index System) [19], Lagos (Nigeria) [20,21], Boston (United States) [22], and in Taranto (Italy) [23]. Finally, to predict house price in Iran, fuzzy regression model was adopted and found to be effective compared to artificial neural networks [24].

The main purpose of the current work is to compare the performance of three optimized predictive models in the context of house price evaluation; each one belongs to a different class of technical tools. The first one is boosting ensemble regression trees that closely resemble human reasoning where decisions are taken following on deductive reasoning. The second one is support vector regression which is based on inductive reasoning to separate data in a hyper-plane. The third one is Gaussian process regression which is a non-parametric method that belongs to advanced statistical models used to approximate shape of functions. Indeed, the origins of these predictive systems are clearly distinct and their respective underlying algorithms differ greatly. In addition, boosting ensemble regression trees and support vector regression are assumptions-free, whilst Gaussian process regression assumes standard statistical assumptions such as stationarity, normality, and independency. Therefore, the findings from this study will enable a better assessment of these different predictive systems in the problem of house price evaluation by means of various performance measures to identify the better one.

In this regard, we attempt to compare our results to those obtained in [16] in their recent interesting study where they found that artificial neural networks outperformed multivariate regression models when applied to a large database. Thus, our contribution is threefold. First, we implement and compare the performance of boosting ensemble regression trees, support vector regression, Gaussian process regression in the task of predicting house price. Indeed, these models have not been employed and validated on the same problem. Second, the parameters of the three models are optimized for better fitting of the data and accurate predictions. In this regard, Bayesian optimization algorithm is employed. Indeed, contrary to the literature, we use optimization for better tuning of the parameters of the models. Third, we use the same large database as in [16] where artificial neural networks outperformed multivariate regression models. Hence, the models we adopt and optimize by Bayesian optimization (boosting ensemble regression trees, support vector regression, Gaussian process regression) will be compared to artificial neural networks and multivariate regression models.

Recall that support vector machines are powerful models successfully employed in various applications; including stock price forecasting [25], cryptocurrency price forecasting prediction [26], cryptocurrency trading volume prediction [27], credit risk evaluation [28], bank telemarketing [29], financial risk forecasting [30], solar radiation prediction [31], beam multi-damage detection [32], and earth-rock dam control [33]. Besides, boosting ensemble decision trees are capable to improve prediction ability of regressors [34] and are found to be effective in a variety of managerial applications [35,36] and scientific problems [37,38]. Finally, Gaussian process regression is an efficient model widely used in engineering problems [39–41].

For optimal tuning, the Bayesian optimization (BO) algorithm [42] is used in the current work to find optimal values of key parameters of boosting ensemble regression trees, support vector regression, and Gaussian process regression. Bring in mind that Bayesian optimization algorithm allows sampling of several thousand points within the variable bounds, takes several of the best feasible points, and improves them using local search in order to find the apparent best feasible point [42]. In addition, the algorithm is fast since the best feasible points depend on the modeled posterior distribution.

The rest of the paper is organized as follows: The three predictive systems (ensemble regression trees, support vector regression, and Gaussian process regression), Bayesian optimization, and experimental design and performance measures are presented in Section 2. The simulation results are presented in Section 3. Section 4 discusses the results. Finally, Section 5 concludes the paper.

2. Methods

2.1. Boosting ensemble regression trees

Regression trees are used to construct predictive models from data by recursively partitioning the data space into subsets and fitting a

predictive model within each subset. Accordingly, data partition is viewed graphically as a decision tree. In the current study, we focus on boosting ensemble systems which are composed of homogeneous sub-systems represented by several regression trees. Specifically, in our work, the least squares boosting algorithm [43] is adopted to generate the ensemble regression trees. Let consider β be regression coefficient vector, $\mathbf{X}\,\beta$ be the predicted value of the response, and r = y - $\mathbf{X}\beta$ be the residuals. Fix the learning rate (shrinkage factor) L > 0, the number of boosting iterations M, initialize $\hat{r}^0 = y$, $\hat{\beta}^0 = 0$, and set k = 0. Then, the least squares boosting algorithm is described as follows:

(1) For $0 \le k \le M$ do the following:

(2) Find the covariate j_k and $\widetilde{u}_{jk}$ as follows:

$$\widetilde{u}_m = \underset{u\in\mathbb{R}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\hat{r}_i^k - x_{im} u \right)^2 \right) for\, m = 1, \ldots, p \tag{1}$$

$$j_k \in \underset{1\le m\le p}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(\hat{r}_i^k - x_{im} \widetilde{u}_m \right)^2 \right) \tag{2}$$

(3) Update the current residuals and regression coefficients for $j \neq j_k$ as follows:

$$\hat{r}^{k+1} \leftarrow \hat{r}^k - L\mathbf{X}_{jk}\widetilde{u}_{jk} \tag{3}$$

$$\hat{\beta}_{jk}^{k+1} \leftarrow \hat{\beta}_{jk}^k + L\widetilde{u}_{jk} \tag{4}$$

$$\hat{\beta}_j^{k+1} \leftarrow \hat{\beta}_j^k \tag{5}$$

In least squares boosting algorithm framework, the number of boosting iterations M and learning rate L together control the training error and the amount of shrinkage. In our study, the number of learning cycles, learning rate, leaf size, number of splits are optimized by Bayesian optimization (BO) framework [42].

2.2. Support vector regression

The support vector regression (SVR) [44] seeks to map low-dimensional non-linear data points to a high-dimensional space by employing a specific kernel function such that the error distance between the data points and the hyperplane is minimized. Let consider dataset $D = (x_i, y_i)$ for i = 1, 2,…,N where x_i is the input data and y_i is the output data. Then, the nonlinear SVR used to predict y (f(x)) is defined as:

$$f(x) = \sum_{i=1}^{N} (a_i - a_i^*) K(x, x_i) + b \tag{6}$$

where b is the bias, $f(x)$ is the forecasted value of y, a and a* are Lagrange multipliers, and $K(x, x_i)$ is a kernel function. Popular kernels used with support vector machine include linear, Gaussian, polynomial, and sigmoid. In our study, Bayesian optimization [42] is employed to find optimal choice of kernel function along with its optimal parameters.

2.3. Gaussian process regression

The Gaussian regression (GR) [45] is a collection of random variables that follows multivariate Gaussian process where the probability distribution function $f(x)$ is defined as follows:

$$f(x) \sim GP\left(m(x), K(x, x')\right) \tag{7}$$

where $m(x)$ is the mean and $k(x, x')$ is covariance (kernel function) which are expressed as follows:

$$m(x) = E\left[f(x)\right] \tag{8}$$

$$K(x, x') = E\left[\left(f(x) - m(x)\right)\left(f(x') - m(x')\right)^T\right] \tag{9}$$

In this framework, the mean (x) is the expected the distribution of x and the covariance matrix (x, x') is used to describe the correlation between each random variable. Both Eqs. (7) and (8) describe a prior distribution over functions by including prior assumptions about $f(x)$. In general, the kernel function $k(x, x')$ approximate $f(x)$ by $f(x')$ given the two points x and x'. In this study, the kernel function used to approximate the covariance matrix $K(x, x')$ is defined as:

$$k(x, x') = exp[-\gamma \|x - x'\|^2] \tag{10}$$

where γ is a constant parameter used to describe the width of the Gaussian kernel. In our work, its optimal value is automatically optimized by employing the Bayesian optimization algorithm [42].

2.4. The Bayesian optimization

The Bayesian optimization (BO) [42] seeks to find the global optimum by incorporating prior belief about the objective function $f(x)$ and updates the prior with observations taken from $f(x)$ to obtain a posterior that improves approximation of $f(x)$. In addition, the Bayesian optimization employs an acquisition function that applies sampling in search sets where an improvement over the current best observation is probable. For instance, let $f(x)$ be the objective function and $EI(x,Q)$ be the expected improvement based on the posterior distribution function Q. then, $EI(x,Q)$ is given by:

$$EI\,(x, Q) = E_Q\left[\max\left(0, \mu_Q\left(x_{best}\right) - f\,(x)\right)\right] \tag{11}$$

where x_{best} is the location of the lowest posterior mean and $\mu_Q(xbest)$ is the lowest value of the posterior mean.

In our study, the BO algorithm is adopted to optimize (*i*) the number of learning cycles, learning rate, leaf size, and number of splits of the boosting ensemble regression trees, (*ii*) to find the optimal choice of kernel function along with its optimal parameters for the SVR, and (*iii*) to determine optimal value of γ for the Gaussian regression. In this regard, the BO technique is employed through k-fold cross validation which is a very popular approach in machine learning algorithms. In this regard, the number of folds is set to ten.

2.5. Protocol of experiments and performance measures

We adopt ten-fold cross-validation method to train and test each predictive model using Matlab 2019a© machine learning toolbox. The performance of each predictive model is evaluated by computing the following usual metrics: the mean absolute error (MAE), root mean squared error (RMSE), mean absolute relative error (MARE), and mean absolute percentage error (MAPE). They are described as follows:

$$MAE = n^{-1} \sum_{i=1}^{n} \left|A_i - P_i\right| \tag{12}$$

$$RMSE = \sqrt{n^{-1} \sum_{i=1}^{n} \left(A_i - P_i\right)^2} \tag{13}$$

$$MARE = n^{-1} \sum_{i=1}^{n} \frac{\left|A_i - P_i\right|}{A_i} \tag{14}$$

$$MAPE = n^{-1} \sum_{i=1}^{n} \left|\frac{A_i - P_i}{A_i}\right| \times 100\% \tag{15}$$

where A and P represent respectively the actual and predicted value, i is instance index, and n is total number of out-of-sample data points.

3. Data and results

The data used to train and test boosting ensemble regression trees, support vector regression, and Gaussian process regression consists of five instances (attributes) and one output representing house price. The five input instances are house age in year, distance to the nearest transportation station in meter, number of convenience stores in the living circle on foot, geographic coordinate in terms of latitude in degree unit, and geographic coordinate in terms of longitude in degree unit. The total number of instances is 414. The data is obtained from

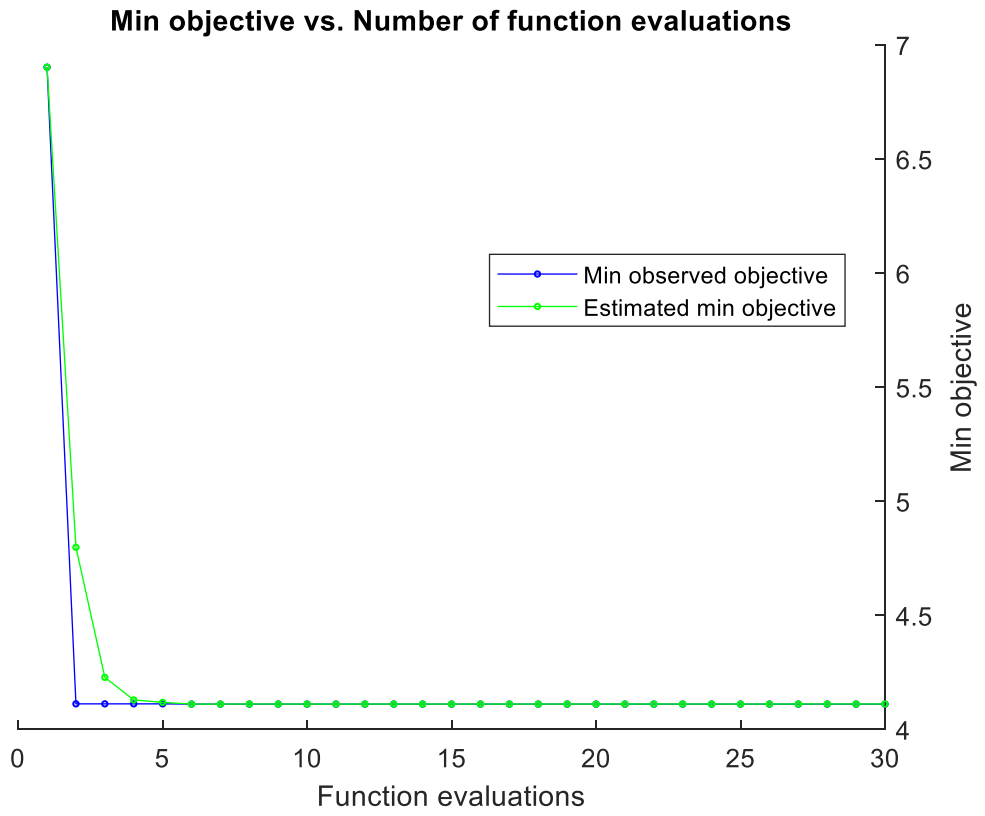

Fig. 1. Plot of minimum objective function depending on function evaluations: boosting ensemble regression trees predictive model.

Taiwan Ministry of the Interior during the period of June 2012 to May 2013 and gathered from two districts in Taipei City and two districts in New Taipei City. The convergence of minimum objective depending on functions evaluation from Bayesian optimization algorithm is displayed in Fig. 1 when applied to boosting ensemble regression trees, Fig. 2 when applied to support vector regression, and Fig. 3 when applied to Gaussian process regression predictive model. Recall that Bayesian Optimization applies a direct the search in order to find the minimum or maximum of an objective function from a Bayesian perspective. Accordingly, the optimal parameters are found by using the predicted mean and predicted variance generated by the normal distribution model. As shown in Figs. 1–3, both minimum observed objective and estimated minimum objective decrease at the same fast rate in a simultaneous manner. In other words, Bayesian optimization is reasonably encouraging and acceptable to tune the predictive systems considered in the current work.

Besides, Fig. 4 shows the boxplot of prediction error associated with each predictive system. It can again be seen that boosting ensemble regression trees have the smallest error rate median and low error variability indicated by the range of the distribution. In addition, support vector regression has the highest error median and Gaussian process regression has the largest error variability. Hence, boosting ensemble regression trees provide stable and low prediction error compared to the other predictive systems. Table 1 summarizes the evaluation results of the aforementioned predictive systems in terms of RMSE, MAE, MARE, and MAPE performance measures. As shown in the table, boosting ensemble regression trees yielded to the lowest performance measures followed by Gaussian process regression and support vector regression respectively. Thus, the overall accuracy of boosting ensemble regression trees is higher than those of Gaussian process regression and support vector regression. Obviously, the comparison of the boosting ensemble regression trees over the other single predictive systems shown in Table 1 is helpful in order to understand whether ensemble predictive systems can outperform the single ones; for instance,

Table 1
Performance metrics.

	ERT	SVR	GPR
RMSE	5.4240	6.4450	6.2214
MAE	3.8032	5.1417	4.7641
MARE	0.1051	0.1441	0.1350
MAPE	1.1049	2.0771	1.8224

Gaussian process regression and support vector regression. Accordingly, we find that the ensemble predictive system; namely the boosting ensemble regression trees, performs better than single best predictive systems when tested on the data set at hand. However, constructing and optimizing ensembles predictive systems such as boosting ensemble regression trees requires larger memory and computational time than constructing a single optimized predictive system; for instance, Gaussian process regression and support vector regression.

4. Discussion

Bear in mind that [16] achieved an average MRSE of 8.04 by using multivariate regression models and 7.12 by using artificial neural networks models; for instance, a multi-layered perceptron, when both validated on the same data set as ours. Since boosting ensemble regression trees, Gaussian process regression, and support vector regression yielded respectively to RMSE value of 5.4240, 6.2214, and 6.4450, they clearly outperformed multivariate regression models and artificial neural networks employed by [16] on the same data set.

The underperformance of artificial neural networks in [16] against boosting ensemble regression trees, Gaussian process regression, and support vector regression could be explained by the fact that the data size is relatively small as the total number of instances is only 414. Indeed, artificial neural networks are data consuming intelligent machines capable to approximate nonlinear functions; but they require

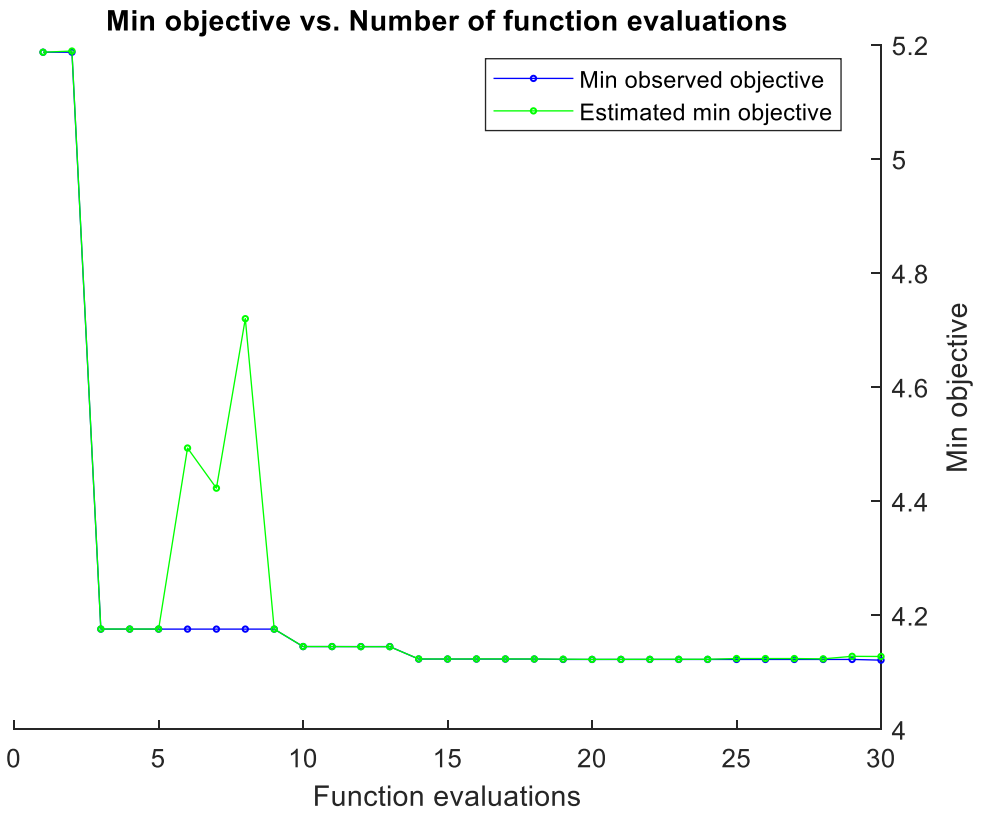

Fig. 2. Plot of minimum objective function depending on function evaluations: support vector regression predictive model.

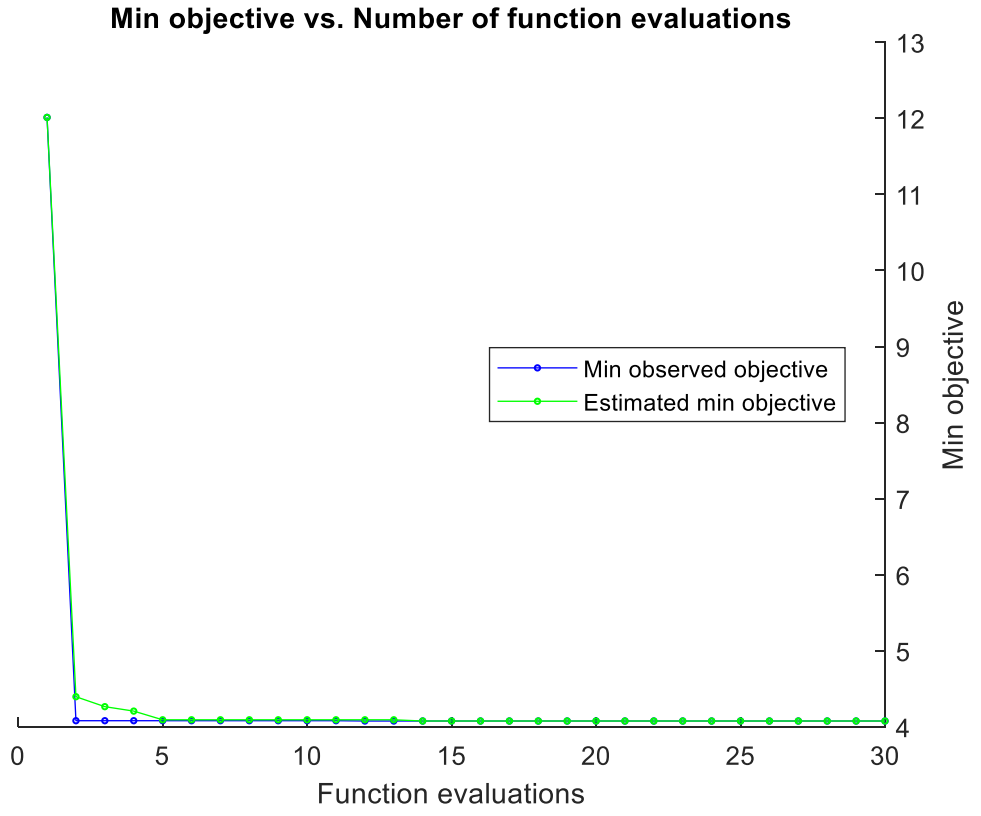

Fig. 3. Plot of minimum objective function depending on function evaluations: Gaussian process regression predictive model.

very large number of examples to efficiently learn data for better function approximation. Besides, multivariate regression models employed in [16] are linear statistical models which are sensitive to normality distribution, nonlinearity, and outliers in data. On the other hand,

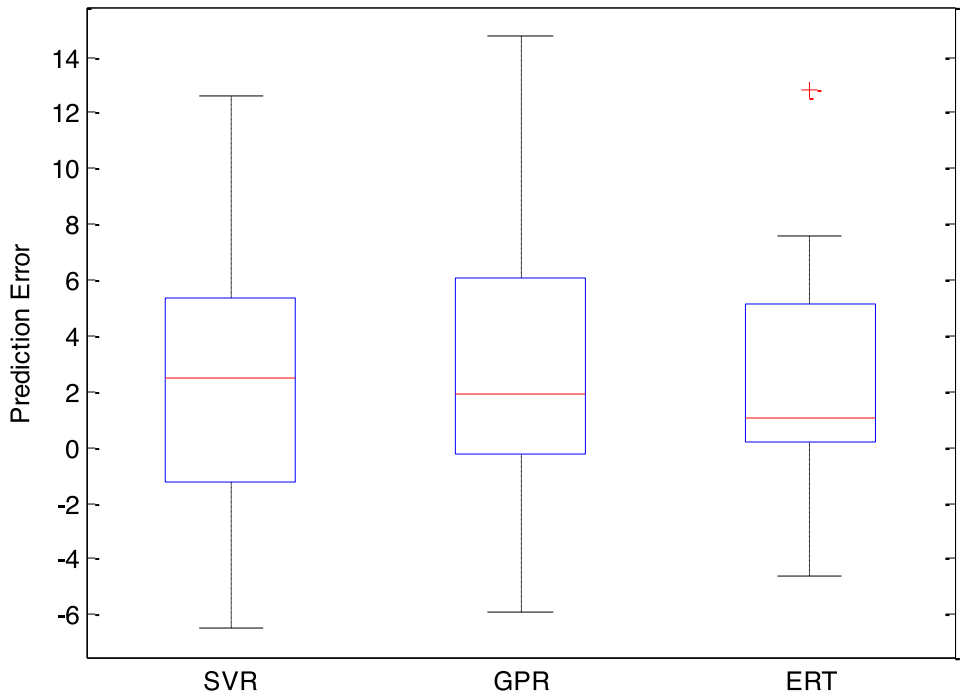

Fig. 4. Boxplot of prediction errors for each predictive system: support vector regression (SVR), Gaussian process regression (GPR), and ensemble regression trees (ERT). The horizontal line within each boxplot indicates the median of the distribution and cross in red font indicates outlier. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

boosting ensemble regression trees are kind of artificial intelligence methods that take advantage of their relative processing simplicity and parallelization technique. Specifically, regression trees determine a set of if-then rules and minimize the error cost by considering both error rate and variance. They do not require assumptions regarding the distribution of predictors and can grip highly skewed numerical data.

One major limit of the current study is that some macro and microeconomic variables have not been considered for house price evaluation. Indeed, various works have shown that housing market depends on general economy factors across countries. For instance, the impact of monetary policy on house prices in South Africa was found to be larger in bear regime than in bull regime [25]. Also, the levels of income, construction costs, impending marriages, user cost and land prices are the primary determinants of house prices in [46], and housing prices in more developed provinces are determined by construction costs and land prices [46]. In addition, the combination of exchange rate regulation and property tax contributes to the stability of the housing market in China [47]. Furthermore, interest rate spreads, real stock market growth, growth in real personal disposable income per capita and inflation are predictors of house prices [48]. Moreover, financial developments in other asset markets can play a significant role as a trigger in the emergence of explosiveness in international housing markets [48].

Unfortunately, the publicly available database we used is limited and does not contain macroeconomic and microeconomic information in Taiwan. Indeed, in general, house pricing data is small and limited [10,12] and imprecise and ambiguous [11]. However, our study has the merit to compare the effectiveness of the presented machine learning techniques in predicting house price when the number of inputs is strictly limited. Indeed, contrary to previous works [25,46–48] where standard linear econometric models were adopted to investigate the relationship between stationary macroeconomic variables and stationary housing price variations, our paper uses and compares advanced machine learning methods to predict nonstationary housing price data with very limited number of nonstationary predictors. In other words, we implemented and compared the performance of three different machine learning methods which is basically a data analytics problem, whilst previous works [25,46] [47,48] are mainly dealing with statistical estimation and inference of the relationship between macroeconomic predictors and the house price. Therefore, our study enriches the limited existing works on predictive analytics of housing price [10–16]. In this regard, we are aware that our study results apply to Taiwanese housing market as we are limited to the access of such database, but the comparison between these three models can also be extended to any other housing markets worldwide. This is left for future work.

Bring in mind that boosting ensemble regression trees and SVR are nonparametric models that are not based on assumptions regarding statistical distributions of the data and specific parametric function forms, contrary to econometric models employed in some previous works [25, 46–48]. In addition, machine learning methods; such as boosting ensemble regression trees and SVR; use regularization principle, which shrinks the influences of redundant or overfitting predictors to zero. In this regard, bagging regression trees account for nonlinear interactions between predictors and are capable to alleviate multicollinearity [49]. Besides, support vector machines family models (including for instance, SVR) are deterministic-learning features machine learning methods which are not sensitive to multicollinearity due to their deterministic solutions of support vectors [49]. Hence, boosting ensemble regression trees and SVR require less data cleaning and are not influenced by outliers and multicollinearity. Finally, Gaussian process regression is not sensitive to multicollinearity since it makes use of a kernel function to compute the approximation function (See Eq. (11), for instance). Indeed, the introduction of a kernel stabilizes the computation of the approximation function which is very effective in presence of noise and multicollinearity. Finally, it is worth to mention that when the goal is to perform a forecasting task (predictive analytics problem as opposition to estimation and inference problem), then multicollinearity is not really a problem under boosting ensemble regression trees, SVR, and Gaussian process regression.

In summary, we used least squares boosting algorithm [49] to construct ensemble regression trees so as to combine weak learners (regression trees) by iteratively focusing in the errors resulting at each step until a suitable strong learner is obtained as a sum of the successive

weak ones. Support vector regression systems are machine learning methods capable to map input vector onto high dimensional feature space by using a nonlinear kernel so that complex problem can be transformed into simpler one. Also, support vector regression is able to achieve global optimum and is efficient even the data sample is small or limited [43,50]. In fact, it is able to conduct learning task with relatively small amount of data [43,50]. Furthermore, in support vector regression framework, the decision making can be made only on few support vectors. Finally, based on using a kernel function to nonlinearly mapping data, Gaussian process regression is flexible and a fully probabilistic predictive system. Besides, our study implemented Bayesian optimization [34] using ten-fold cross validation technique to choose optimal parameter values and the kernel of support vector regression predictive model. The Bayesian optimization was also adopted to find optimal parameter values and structure of boosting ensemble regression trees and Gaussian process regression. To validate the prediction accuracy of the three predictive models, a comparison in terms of various performance metrics has been conducted.

Accordingly, the results of empirical analyses showed that the boosting ensemble regression trees tuned by Bayesian optimization perform the best. Indeed, it outperformed Gaussian process regression and support vector regression predictive models all tuned by Bayesian optimization. In other words, boosting ensemble regression trees which belong to artificial intelligence methods outperformed support vector regression which belongs to machine learning family and Gaussian process regression which belongs to statistical models. Furthermore, boosting ensemble regression trees provide stable and low prediction error. Moreover, all three predictive systems employed in the current work outperformed artificial neural networks and multivariate regression models used in [16] and tested on the same data set.

5. Conclusion

This study employed and compared three predictive systems for the first time namely boosting ensemble regression trees which belong to artificial intelligence methods, support vector regression which belongs to machine learning family and Gaussian process regression which belongs to statistical models; all optimized by Bayesian optimization; to the problem of house price prediction. Based on four different performance measures, the experimental results show that the boosting ensemble regression trees are accurate and reasonable for use in house price evaluation as it outperformed support vector regression and Gaussian process regression. In addition, boosting ensemble regression trees provide stable and low prediction error. Besides, all three predictive systems performed much better than artificial neural networks and multi-variate regression model which were employed in a recent work on the same data set. Certainly, the stability and algorithmic efficiency of boosting ensemble regression trees make them an ideal candidate for house price forecasting when applied to a small data sample with few predictors.

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.

Data availability

Data will be made available on request.

References

[1] A.C. Goodman, T.G. Thibodeau, Housing market segmentation and hedonic prediction accuracy, J. Hous. Econ. 12 (2003) 181–201.
[2] O. Bin, A prediction comparison of housing sales prices by parametric versus semi-parametric regressions, J. Hous. Econ. 13 (2004) 68–84.
[3] R. Gupta, A. Kabundi, S.M. Miller, Forecasting the US real house price index: Structural and non-structural models with and without fundamentals, Econ. Model. 28 (2011) 2013–2021.
[4] P.-F. Chen, M.-S. Chien, C.-C. Lee, Dynamic modeling of regional house price diffusion in Taiwan, J. Hous. Econ. 20 (2011) 315–332.
[5] T. Kato, Prediction in the lognormal regression model with spatial error dependence, J. Hous. Econ. 21 (2012) 66–76.
[6] R. Kouwenberg, R. Zwinkels, Forecasting the US housing market, Int. J. Forecast. 30 (2014) 415–425.
[7] O. Kettani, M. Oral, Designing and implementing a real estate appraisal system: The case of Québec Province, Canada, Socio-Econ. Plan. Sci. 49 (2015) 1–9.
[8] L. Bork, S.V. Møller, Forecasting house prices in the 50 states using dynamic model averaging and dynamic model selection, Int. J. Forecast. 31 (2015) 63–78.
[9] Y. Wei, Y. Cao, Forecasting house prices using dynamic model averaging approach: Evidence from China, Econ. Model. 61 (2017) 147–155.
[10] H. Kusşan, O. Aytekin, I. Özdemir, The use of fuzzy logic in predicting house selling price, Expert Syst. Appl. 37 (2010) 1808–1813.
[11] A. Azadeh, B. Ziaei, M. Moghaddam, A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations, Expert Syst. Appl. 39 (2012) 298–315.
[12] L.H. Gerek, House selling price assessment using two different adaptive neuro-fuzzy techniques, Autom. Constr. 41 (2014) 33–39.
[13] J. Wang X. Wen, Y. Zhang, Y. Wang, Real estate price forecasting based on SVM optimized by PSO, Optik 125 (2014) 1439–1443.
[14] B. Park, J.K. Bae, Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data, Exp. Syst. Appl. 42 (2015) 2928–2934.
[15] V. Plakandaras, R. Gupta, P. Gogas, T. Papadimitriou, Forecasting the U.S. real house price index, Econ. Model. 45 (2015) 259–267.
[16] I.-C. Yeh, T.-K. Hsu, Building real estate valuation models with comparative approach through case-based reasoning, Appl. Soft Comput. 65 (2018) 260–271.
[17] A.B. Adetunji, O.N. Akande, F.A. Ajala, O. Oyewo, Y.F. Akande, G. Oluwadara, House price prediction using random forest machine learning technique, Procedia Comput. Sci. 199 (2022) 806–813.
[18] A. Soltani, M. Heydari, F. Aghaei, C.J. Pettit, Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms, Cities 131 (2022) 103941.
[19] X. Xu, Y. Zhang, House price forecasting with neural networks, Intell. Syst. Appl. 12 (2021) 200052.
[20] R.B. Abidoye, A.P. Chan, Modelling property values in Nigeria using artificial neural network, J. Prop. Res. 34 (2017) 36–53.
[21] R.B. Abidoye, A.P. Chan, Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network, Pac. Rim Prop. Res. J. 24 (2018) 71–83.
[22] A. Al Bataineh, D. Kaur, A comparative study of different curve fitting algorithms in artificial neural network using housing dataset, in: Naecon 2018 IEEE National Aerospace and Electronics Conference, 2018, pp. 174–178.
[23] V. Chiarazzo, L. Caggiani, M. Marinelli, M. Ottomanelli, A neural network based model for real estate price estimation considering environmental quality of property location, Transp. Res. Procedia 3 (2014) 810–817.
[24] A. Azadeh, M. Sheikhalishahi, A. Boostani, A flexible neuro-fuzzy approach for improvement of seasonal housing price estimation in uncertain and non-linear environments, South Afr. J. Econ. 82 (2014) 567–582.
[25] S. Lahmiri, Minute-ahead stock price forecasting based on singular spectrum analysis and support vector regression, Appl. Math. Comput. 320 (2018) 444–451.
[26] S. Lahmiri, S. Bekiros, Intelligent forecasting with machine learning trading systems in chaotic intraday Bitcoin market, Chaos Solitons Fractals 133 (2020) Paper ID: 109641.
[27] S. Lahmiri, S. Bekiros, F. Bezzina, Complexity analysis and forecasting of variations in cryptocurrency trading volume with support vector regression tuned by Bayesian optimization under different kernels: An empirical comparison from a large dataset, Expert Syst. Appl. 209 (2022) 118349.
[28] S. Lahmiri, Features selection, data mining and finacial risk classification: A comparative study, Intell. Syst. Account. Finance Manag. 23 (2016) 265–275.
[29] S. Lahmiri, A two-step system for direct bank telemarketing outcome classification, Intell. Syst. Account. Finance Manag. 24 (2017) 9–55.
[30] J. Sun, Integration of random sample selection, support vector machines and ensembles for financial risk forecasting with an empirical analysis on the necessity of feature selection, Intell. Syst. Account. Finance Manag. 19 (2012) 229–246.
[31] Z. Ramedani, M. Omid, A. Keyhani, B. Khoshnevisan, H. Saboohi, A comparative study between fuzzy linear regression and support vector regression for global solar radiation prediction in Iran, Sol. Energy 109 (2014) 135–143.

[32] J. Xiang, M. Liang, Y. He, Experimental investigation of frequency-based multi-damage detection for beams using support vector regression, Eng. Fract. Mech. 131 (2014) 257–268.
[33] J. Wang, D. Zhong, H. Adeli, D. Wang, Liu M., Smart bacteria-foraging algorithm-based customized kernel support vector regression and enhanced probabilistic neural network for compaction quality assessment and control of earth-rock dam, Expert Syst. 35 (2018) e12357.
[34] A. Özçift, Forward stage-wise ensemble regression algorithm to improve base regressors prediction ability: An empirical study, Expert Syst. 31 (2012) 1–8.
[35] S. Figini, R. Savona, M. Vezzoli, Corporate default prediction model averaging: A normative linear pooling approach, Intell. Syst. Account. Finance Manag. 23 (2016) 6–20.
[36] W.-C. Lin, Y.-H. Lu, C.-F. Tsai, Feature selection in single and ensemble learning-based bankruptcy prediction models, Expert Syst. 36 (2019) e12335.
[37] N. Gupta, N. Ahuja, S. Malhotra, A. Bala, G. Kaur, Intelligent heart disease prediction in cloud environment through ensembling, Expert Syst. 34 (2017) e12207.
[38] M. Naghizadeh, N. Habibi, A model to predict the survivability of cancer comorbidity through ensemble learning approach, Expert Syst. 36 (2019) e12392.
[39] Y.-J. He, J.-N. Shen, J.-F. Shen, Z.-F. Ma, State of health estimation of lithium-ion batteries: A multiscale Gaussian process regression modeling approach, Am. Inst. Chem. Eng. 61 (2015) 1589–1600.
[40] Y.-J. He, Z.-F. Ma, A data-driven Gaussian process regression model for two-chamber microbial fuel cells, Fuel Cells Fundam. Syst. 16 (2016) 365–376.
[41] Y. Okadome, Y. Nakamura, H. Ishiguro, Sampling-based motion planning with a prediction model using fast Gaussian process regression, Electron. Commun. Japan 100 (2017) 24–34.
[42] M. Gelbart, J. Snoek, R.P. Adams, Bayesian optimization with unknown constraints, 2014, https://arxiv.org/abs/1403.5607.
[43] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Statist. 29 (2001) 1189–1232.
[44] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[45] C.E. Rasmussen, C.K. Williams, Gaussian Processes for Machine Learning, Massachusetts Institute of Technology: MIT-Press, 2006.
[46] Q. Li, S. Ch, House prices and market fundamentals in urban China, Habitat Int. 40 (2013) 148–153.
[47] Y. He, F. Xia, Heterogeneous traders, house prices and healthy urban housing market: A DSGE model based on behavioral economics, Habitat Int. 96 (2020) 102085.
[48] E. Martínez-García, V. Grossman, Explosive dynamics in house prices? An exploration of financial market spillovers in housing markets around the world, J. Int. Money Finance 101 (2020) 102103.
[49] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[50] V. Vapnik, S. Golowich, A. Smola, Support vector machine for function approximation, regression estimation, and signal processing, Adv. Neural Inf. Process. Syst. 9 (1996) 281–287.