Contents lists available at ScienceDirect

Habitat International

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

An explainable model for the mass appraisal of residences: The application of tree-based Machine Learning algorithms and interpretation of value determinants

Muzaffer Can Iban

Department of Geomatics Engineering, Mersin University, Çiftlikköy Campus, 33343, Mersin, Türkiye

ARTICLE INFO

Keywords:
Machine learning
Explainable artificial intelligence
Real estate mass appraisal
Feature importance
Interpretability

ABSTRACT

In the mass appraisal of properties, Machine Learning (ML) algorithms have produced effective and promising results. Analysts use various algorithms to train their models with limited data and make appraisals on large data sets. However, research into which value determinants the models take into account when appraising values is insufficient. This research looks at how eXplainable Artificial Intelligence (XAI) methods can be integrated with mass real estate appraisal studies. Experimental studies were carried out on a data set containing 1002 samples and 43 independent variables. Tree-based ML regressors, namely Random Forest, XGBoost, LightGBM, and Gradient Boosting, were used for training the predictive models. The performance of these regressors was compared with that of classical multiple regression analysis. The Permutation Feature Importance (PFI) technique was used for the selection of the variables that contributed the most to the training of the models. Models retrained with selected variables were locally interpreted using the SHapley Additive eXplanations (SHAP) method. In this way, it was possible to observe the value determinants that contribute to the price estimation of each real estate sample. This study demonstrates that XAI approaches should be integrated into mass real estate valuation systems specifically, and into urban and housing research more generally, helping analysts and scholars to explain their models more transparently. The outcomes of this study can be a harbinger for analysts and scholars who wish to explain their models more transparently. Last but not least, this study advocates the use of tree-based ML algorithms since they not only allow us to implement XAI approaches but also outperform the stand-alone ML regressors.

1. Introduction

Land and other types of immovable property are typically taxed based on their values in well-established market economies (Gnat, 2021). A Mass Appraisal System (MAS) of real properties is a general name for a collection of techniques used to appraise a large number of properties uniformly and quickly. An accurate and up-to-date MAS is an essential tenet of fair property assessment for tax purposes (Hefferan & Boyd, 2010), and while each nation currently tends to have its own method, future advancements in global real estate markets may result in comparable MAS procedures. Good practices bring about a healthy real estate market, including stable tax policies, easy access to credit, and mortgage options (Filippakopoulou & Potsiou, 2014). Furthermore, an MAS helps planners and investors to understand the trends and preferences after implementing land use decisions or new investments (Hei-Ling Lam & Chi-Man Hui, 2018; Li, Hui, & Shen, 2020; Ling & Hui, 2013; Ming & Hin, 2006).

Real estate appraisal studies rely on the prices of a few recently sold properties nearby as comparable evidence due to the incomplete knowledge about the current market pricing. Hence, traders use local market knowledge to determine the open market value of the real estate as well as their own projected values. In actuality, traders could also employ a variety of other information sources, such as expert reports, deed transfers, or even valuers' own expert predictions regarding the direction of the market (Doumpos, Papastamos, Andritsos, & Zopounidis, 2021; Kok, Koponen, & Martínez-Barbosa, 2017; Lo, Chau, Wong, McCord, & Haran, 2022). Because market values are heavily influenced by the physical and locational characteristics of properties (McCluskey, Deddis, Mannis, McBurney, & Borst, 1997; Yang & Bai, 2013), property values can be predicted more quickly, more accurately, and with greater contextual specificity thanks to more and better data and developing computing capabilities (Tajani, Morano, & Ntalianis, 2018; Glumac &

E-mail address: caniban@mersin.edu.tr.

https://doi.org/10.1016/j.habitatint.2022.102660
Received 5 May 2022; Received in revised form 16 August 2022; Accepted 24 August 2022
Available online 31 August 2022

des Rosiers, 2021; Gnat, 2021; Leao et al., 2021).

Increased data availability, along with ML advancements, has resulted in various applications across industries (Alpaydin, 2020; Bishop, 2006) with scholars now able to extract valuable information from, and uncover hidden patterns in much larger databases (Grekousis, 2019). Many recent publications are built on the use of ML models to tackle research questions arising in urban (Chaturvedi & de Vries, 2021; Jia, Zhang, Huang, & Luan, 2022; Kaczmarek, Iwaniak, Świetlicka, Piwowarczyk, & Nadolny, 2022; Resch & Szell, 2019; Xu et al., 2022) and housing studies (Fields & Rogers, 2021; Amarasinghe Arachchige et al., 2022). In the realm of MAS, ML-based studies have typically focused on the prediction performance of the algorithms, but less attention has been paid to the explainability of these models and this has important consequences for their interpretability, robustness and, ultimately, trustworthiness. This work, therefore, sits at the nexus of the explainability of ML approaches and MAS studies, showing a path for an explainable MAS concept, which is the main motivation of this article.

2. Literature review and scope: towards an explainable MAS concept

Individual appraisal approaches are commonly used for single properties. Nevertheless, it becomes more challenging and time-consuming when many properties are appraised individually. As Grover stated (2016), when a substantial number of properties must be appraised, an MAS should be used by employing standardized processes, data, attributes, and algorithms. Appraisal values may also be used for occasional taxes or to verify taxpayers' self-assessments. However, setting up an MAS has challenging processes. Input data sets should be sufficiently reliable. There must be a viable and open real estate market where transactions for all sorts of property can be done easily with a medium through which buyers, sellers, appraisers, surveyors, and estate agents can learn about the market conditions (Bartke & Schwarze, 2021). In developing countries, such open markets are hard to find, and this situation creates difficulties in the creation of large data sets needed to establish a robust MAS (Grover, 2016). Several studies using various models addressed this issue and uncovered more accurate results using very limited data sets, and scholars are developing new methodologies and making insightful contributions to the existing body of literature.

Hass' (1922) Hedonic Price Model, a.k.a. Multiple Regression Analysis (MRA), has been the dominant regression method in appraisal studies. The sale price is the dependent variable and the property features are the independent variables which reflect the consumer preferences (Colwell & Dilmore, 1999). Theoretically, the MRA is a function of the locational and physical characteristics of real estate (Hamilton & Morgan, 2010). Although the MRA is widely acknowledged and regarded as an orthodox method in MAS studies (Bunyan Unel & Yalpir, 2019; Sisman & Aydinoglu, 2022), multicollinearity between feature variables and outlier samples can seriously impair the performance (Bilgilioğlu & Yılmaz, 2021). Furthermore, some authors claim that MRA may be so simple that it can generate biased or underestimated predictions (Hui, Chau, Pun, & Law, 2007; Selim, 2009; Suparman, Folmer, & Oud, 2014; Wang & Li, 2019), particularly when the data patterns exhibit non-linearity (Connellan & James, 1998; Lenk, Worzala, & Silva, 1997). Machine Learning (ML) algorithms have been proposed as possible solutions to handle linear and nonlinear relationships in a dataset where both categorical and numerical variables exist. A definition of supervised ML is given by Watson (2019):

> *"The typical supervised learning setup involves a matrix of features X (predictors, independent variables, etc.) and a vector of outcomes Y (the response, dependent variable, etc.) that together form some fixed but unknown joint distribution P (X, Y). The goal is to infer a function f that predicts Y based on X. A model f is judged by its ability to generalize, i.e., to successfully predict outcomes on data that were not included in its training set"*.

Scholars have studied different ML algorithms, focusing on their predictive power in MAS. The size of the data set in each of these studies varies. Some researchers have worked with big data sets (Alfaro-Navarro et al., 2020; Tchuente & Nyawa, 2022), while others have dealt with limited ones (Antipov & Pokryshevskaya, 2012; Aydinoglu, Bovkir, & Colkesen, 2021; Yilmazer & Kocaman, 2020). The MRA based on Ordinary Least Squares (OLS) has been the most popular one in the existing body of literature (McCluskey, McCord, Davis, Haran, & McIlhatton, 2013). As aforementioned, the MRA has been reported as having some drawbacks. Therefore, the researchers started to focus on other types of regression algorithms (Wang & Li, 2019), namely the Support Vector Machines (SVM) (Chen, Ong, Zheng, & Hsu, 2017; Ho, Tang, & Wong, 2021), Artificial Neural Networks (ANN) (Bilgilioğlu & Yılmaz, 2021; Kathmann, 1993; Tchuente & Nyawa, 2022), and logistic regression. These algorithms are suitable for estimating the link between price and numerous features. Nevertheless, the relationship between them is mostly unclear (Selim, 2009). Therefore, these approaches are not explainable, and often they are considered black-box models.

It is not sufficient to demonstrate that the algorithm makes accurate predictions, for reasons of transparency and fairness, it should be possible to attribute the role of the input data features in the prediction process. This is called "the explainability of the models," which enables the analyst to determine the most important features contributing to the real estate value. An explainable MAS concept will be useful in getting rid of the problems encountered with feature selection methods that are frequently used in the literature: for example, market conditions may vary greatly between locations, necessitating multiple surveys (Yalpir, Sisman, Akar, & Unel, 2021), and some studies prefer to select features prior to training their models by using only correlation coefficients (Pai & Wang, 2020; Yilmazer & Kocaman, 2020) or multicollinearity analysis (Chen et al., 2017). Nevertheless, these approaches do not give a clue about the contribution of feature variables to predicting the price of a property.

We owe the explanation of the models to the growing popularity of Decision Tree (DT) based ML algorithms. The Random Forest (RF) algorithm aggregates multiple DTs in parallel (Alfaro-Navarro et al., 2020; Aydinoglu et al., 2021; Bilgilioğlu & Yılmaz, 2021; Ho et al., 2021; Yilmazer & Kocaman, 2020; Čeh, Kilibarda, Lisec, & Bajat, 2018). That is why the RF is an example of a *bagging* algorithm. Some other tree-based algorithms build multiple DTs to learn sequentially, and they are known as *boosting* algorithms, such as Gradient Boosting (GB) and Adaptive Boosting (AdaBoost). They have been studied by a small number of researchers (Fedorov & Petrichenko, 2020; McCluskey, Zulkarnain Daud, & Kamarudin, 2014). Since these bagging and boosting algorithms use multiple DTs throughout the sequence of learning, they are known as *ensemble methods* (Alpaydin, 2020). These algorithms work well in classification and regression, with high accuracy, stability, and interoperability (Wang & Li, 2019). For the first time, Antipov and Pokryshevskaya (2012) applied the RF learner in MAS and discovered that it outperforms other models. Subsequent contributions also suggest that RF is a promising learner in MAS studies, and this brings about a need for the comparison of RF with other tree-based ensemble learners. Furthermore, tree-based ensemble models are regarded to be more explainable than neural networks and kernel techniques, in addition to their good performance and simplicity. The increased use of tree-based models in MAS studies has increased efforts to present models in an explainable way. This is the point where the MAS studies meet the subfield of eXplainable Artificial Intelligence (XAI).

The XAI discusses the explainability of ML and deep learning models. The explainability of the models illustrates the contribution of input variables to the overall model predictions, enabling the analyst to identify what the model takes into account when estimating real estate prices (Das & Rad, 2020; Islam, Eberle, Ghafoor, & Ahmed, 2021; Konstantinov & Utkin, 2021; Samek, 2020; van der Waa, Nieuwburg, Cremers, & Neerincx, 2021). The XAI approaches can be split into two broad categories based on their scope: global and local approaches. The trained model is referred to in global explanations. The critical variables to the model and their effects on an average prediction are given in a global explanation of the model (Delgado-Panadero, Hernández-Lorca, García-Ordás, & Benítez-Andrades, 2022). Permutation Feature Importance (PFI) estimates the contribution of each input variable to the prediction of complicated ML models. PFI ignores unimportant features during permuting and keeps the model error constant (Adadi & Berrada, 2018) and can be used to select the most important features, omit the least important ones, and re-train the regressors (Huang, Lu, & Xu, 2016). Certainly, there are other sorts of global explanation approaches. Default feature importance (based on 'Gini importance', a.k.a. mean decrease impurity), or drop-column importance have been implemented in several appraisal studies for feature selection purposes (Hong, Choi, & Kim, 2020; Taecharungroj, 2021). Nevertheless, the default feature importance approach is unreliable when input variables vary in a number of categories, data types, and scales (Strobl, Boulesteix, Zeileis, & Hothorn, 2007). Furthermore, reshuffling a column is more practicable than dropping it since reshuffling retains the distribution of feature variables. As a result, the greater the importance of the variable, the more divergent the two sets of predictions will be (Cascarino, Moscatelli, & Parlapiano, 2022).

The other XAI category is local explainability, which enables the analyst to understand which feature variables contribute more to any of the predictions, and these variables affect the single prediction (Delgado-Panadero et al., 2022). Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are emerging as state-of-the-art local approaches. For each input sample to be explained, the LIME generates new and similar data samples. The model is re-trained, and the LIME compares the differences between predictions generated by the original and randomly generated data samples. The main limitation of the LIME is the locality around a data sample. The other data samples, collected in the real world, may not be similar to the input data (Doumard et al., 2022; Ghalebikesabi, Ter-Minassian, DiazOrdaz, & Holmes, 2021). The main idea of SHAP approach is to see what happens in the model when a feature is missing. This avoids re-training the complex model without generating new values for the feature of interest, and without losing the real-world reality (Amparore, Perotti, & Bajardi, 2021). Lundberg et al. (2020) developed the *TreeExplainer* library which enables the analyst to explain each decision of a tree-based algorithm in terms of the contribution to the prediction. The *TreeExplainer* package uses SHAP values as a richer type of local explanation. The library enables users to prepare SHAP summary plots, which briefly illustrate the extent, frequency, and direction of the contribution of a feature variable.

Using data collected in the Yenişehir district of Mersin Province, Türkiye, this research puts forward three innovations:

1) In addition to tree-based algorithms using bagging methods like RF, novel tree-based ensemble boosting techniques like Gradient Boosting (GB), eXtreme Gradient Boosting (XGBoost), and Light Gradient Boosting (LightGBM) machines are evaluated in this research and their performances are compared with the OLS, as a baseline.
2) This study adopts PFI approach to define the global importance of all input variables and select the most contributing ones for training the selected regressors. Using the feature importance scores resulting from PFI computation, the features that have a weight score greater than 0.01 are selected to train the final regressors.
3) Then, the predictions of the regressors were subjected to a local explanation with the help of SHAP technique. In the Results and Discussions section, the factors affecting the residential prices obtained for Yenişehir with the help of a local explanation will be presented in comparison with the findings obtained by other scholars. The geo-data frame used in this study involves 43 independent variables adopted from previous studies, namely physical, locational, and topographical features of 1002 residential units in the Yenişehir District. To evaluate the performance of these regressors, several metrics recommended by international real estate appraisal guidelines are used. Lastly, the predictions generated by each model are subjected to a pairwise significance test, namely, the Wilcoxon signed-rank test, to understand the statistical similarities and differences between the predictions of different models.

The findings of this study will aid real estate appraisers, as well as policymakers and urban planners, in improving the prediction performance of mass real estate appraisal systems using complex data sets and in interpreting the trends and preferences in real estate values in urban areas. It is also hoped that the techniques elaborated here will also see wider adoption in the urban studies community more generally where ML-based methods are beginning to feature more heavily.

3. Materials and methods

3.1. Study area

Yenişehir is located in the Mersin Metropolitan Area, near the Mediterranean Sea, at 36° 47′ N, 34° 36′ E. Yenişehir's population was 274,944 in 2021, according to census data. Yenişehir is home to around 25% of Mersin's population. The population growth rate is greater than the Mersin average, at around 3.7 percent each year. Urbanization began in the 1980s and accelerated in the 2000s with the development of high-rise apartments. While the areas close to the coastline and those near the port have older building stock, the northern and western areas of the district have a modern and dense construction texture. In the district, there are four shopping malls, three university campuses, a highway to other metropolitan areas, sports and recreation facilities, and cultural centers. The Yenişehir district serves as a gathering place for Mersin residents as the city's new heart. As a result, the housing stock in the Yenişehir district has grown appealing and has become a high-value housing production hub due to the aging of the building stock in the rest of the city. Because Mersin does not have a pre-assumed earthquake risk, there is no major height limit for the structures. The city of Mersin, on the other hand, has a hot and humid Mediterranean climate that favors detached dwellings.

3.2. Price data

The data on the cadastral infrastructure and land management systems in Türkiye does not yet store the market values of real estate. In buying and selling transactions, the values declared by citizens do not reflect the market realities and are usually kept low for taxation and fee calculation. Therefore, in real estate valuation studies in Türkiye, data scraped from files kept by real estate appraisers or from web pages that publish real estate advertisements is used (Çağdaş, 2013).

The sales prices, locations, and physical properties of 1002 residential flats were manually gathered between January and June 2021 as a result of telephone and face-to-face interviews with real estate firms and certified appraisers in the Yenişehir district. The sales prices for the NUTS2 regions of Adana and Mersin were adjusted using the residential price index published by The Central Bank of Türkiye. According to this index, sales prices increased 1.8% in January and March, 2.7% in February and May, and 4.7% in April 2021. As a result, the prices were adjusted to reflect the monthly inflation rate.

3.3. Geospatial data frame

Physical features determine the fabric of each residence, whereas location-based features determine the accessibility to local services and interests (Bin, Gardiner, Li, & Liu, 2020). Since no preliminary analysis has been made regarding the purchasing preferences of the residents in Yenişehir and the dynamics of the internal market of Mersin City are unknown, the list of these features needs to be kept wide. A feature selection procedure is a must to highlight the most important value determinants and omit the unnecessary ones. The features determine the preferences and behaviors of the buyers, and these features are specific to the cities, even neighborhoods. Physical features are very important to buyers almost everywhere in the world (Kang, Zhang, Peng, et al., 2021; Xiao, Hui, & Wen, 2019). Furthermore, there are some studies reporting that houses close to urban services and green areas positively affect the purchasing behavior of potential customers (Daams, Sijtsma, & Veneri, 2019; Giannico et al., 2021; Su et al., 2021; Yuan, Wei, & Wu, 2020). Social experiences and demographic characteristics are also important indicators that define preferences (Kang, Zhang, Gao, Peng, & Ratti, 2021; Wang, Hui, & Sun, 2017). Last but not least, air quality, urban noise, and proximity to waterscapes are other variables that affect house prices (Hui et al., 2007).

Physical features of the samples were added to the data frame manually. Location-based and topography-based features were generated in the QGIS 3.16 environment for each sample through manual digitization, satellite images, or municipal data. Table 1 shows the data collected from real estate offices as well as the spatial data generated. In the training task, The *Sales Price* feature is the dependent feature, while the other 43 features are independent. In the end, a geospatial data frame was obtained as seen in Fig. 1.

3.4. Tree-based ensemble algorithms

In this study, the tree-based ensemble ML algorithms are adopted, and the predictive performance of these algorithms is compared with the performance of so-called OLS. An *ensemble* is a grouping of numerous trained predictors whose combined prediction is made to increase the predictive capacity of a single predictor. Ensemble algorithms can generally deal with noisy data and small outliers without affecting overall performance (González, García, Ser, Rokach, & Herrera, 2020). The architecture of an ensemble can be in two different ways: 1) If the algorithm employs a sequence of learners on the same training samples and repeatedly increases the algorithm's performance, this is called *boosting*. Each successive learners look for the failed predictions from the previous learners and adjust themselves so that the next learner does not make the same failed prediction again. The most popular examples of tree-based ensemble boosting algorithms are GB, XGBoost, and LightGBM. 2) If the algorithm produces different learners from a training sample that was picked at random, and generates a collection of these learners, this is called *bagging*. This time, an average of all predictions from many learners is used, which is more reliable than a single learner. The so-called bagging algorithm is the RF, which has been predominantly used for regression and classification problems in academia (Dietterich, 2000). Specifically, tree-based bagging methods create a forest of different DTs and choose the best one in terms of predictive performance, while tree-based boosting methods create a sequence of DTs until it reaches the best one. In this research, regression tasks with selected algorithms are conducted by using Python's cutting-edge Scikit-learn (Pedregosa et al., 2011), xgboost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017) packages.

The RF, developed by Breiman (2001), is a DT variant that extrapolates the model by training alternative combinations of the data on different features. This yields N separate DTs, which RF then combines into a single model. The training process is started by selecting a random subset $S^{'}$ and the training data is iteratively divided into left (S_l) and right subsets (S_r) by using a predefined threshold and split function. The trees can grow to the maximum depth possible. A majority vote among the assessments provided by the multiple DTs generates the overall prediction of a data sample.

The GB, developed by Friedman (2001), tries to fit a new DT to the residuals of the previous DT. In earlier iterations, GB used a stage process method that relied on the gradient descent of the loss function. After constructing the tree, the terminal leaves are labeled, and their outputs are calculated to reduce the loss function for all inputs in each leaf. These results are weighted by a quantity known as the learning rate. The learning rate modifies the input from new trees, allowing the algorithm to gradually improve its performance. This method is useful when dealing with minimal bias and high variation (González et al., 2020). The loss function can indeed be thought of as the degree of error in the model. The greater the loss function, in general, the more probable the

Table 1
Feature variables in the geospatial data frame.

Category	Feature	Data Type	Source
Price	Sales Price (Turkish Liras)	Float	Real Estate and Certified Appraisal Offices
Administration	Neighborhood	Categorical	
Physical Features	Gross Area (m^2)	Float	
	Net Area (m^2)	Float	
	Floor Number	Integer	
	Building Floors	Integer	
	Number of Bathrooms	Integer	
	Heating System	Categorical	
	Number of Rooms	Integer	
	Site Complex or Not	Boolean	
	Building Age	Integer	
	Balcony	Boolean	
	Elevator	Boolean	
	Number of Facades	Integer	
	Sea Panorama	Boolean	
	Detached or not	Boolean	
Topographical Features	Elevation	Float	ALOS PALSAR Digital Elevation Model of 12.5 m Resolution
	Aspect	Categorical	
	Slope	Float	
Location Features (in km)	Distance to Main Roads	Float	Vector files (Municipal Data)
	Distance to Bus Stops	Float	
	Distance to Waste Disposal Areas	Float	
	Distance Potential Light Rail System	Float	
	Distance to Gas Stations	Float	
	Distance to Schools	Float	
	Distance to Central Business District	Float	
	Distance to Government Buildings	Float	
	Distance to Hospitals	Float	
	Distance to Sport Facilities	Float	
	Distance to Trade Centers	Float	
	Distance to Prayer Halls	Float	
	Distance to Military Zones	Float	
	Distance to Cemeteries	Float	
	Distance to Parks	Float	
	Distance to Industrial Areas	Float	
	Distance to Bazaars	Float	
	Distance to Port	Float	Manual Digitization
	Distance to Universities	Float	
	Distance to Coach Station	Float	
	Distance to Railway Station	Float	
	Distance to Coastline	Float	
	Distance to Shopping Malls	Float	
	Distance to Fair	Float	
	Distance to Highway	Float	

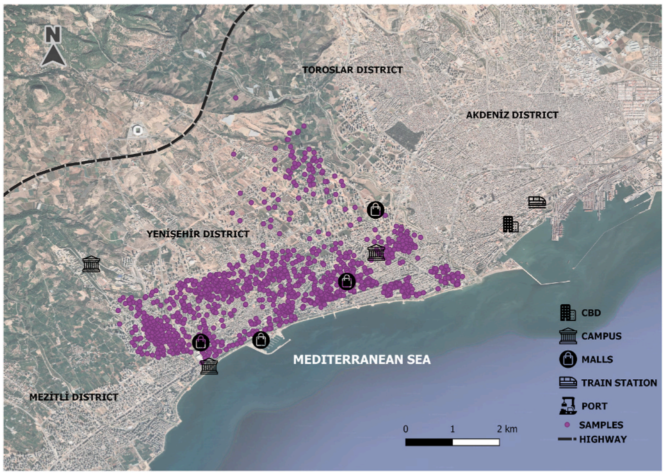

Fig. 1. The geospatial data frame containing the locations of collected samples.

prediction is to fail. The optimum strategy is to bring the loss function to descend in the gradient direction, as the aim is to minimize the loss function.

Chen and Guestrin (2016) presented the XGBoost extending the concept of GB, with the main difference being loss function standardization. The GB algorithm addresses the first derivatives of the loss function, whereas the XGBoost approach employs Taylor's expansion to strengthen the loss function. In other words, GB stops pruning the trees once it encounters a negative loss value, while XGBoost splits the tree until the specified maximum tree depth and makes a backward pruning. While XGBoost uses a level-wise tree growth in which the algorithm seeks the best tree depth, the LightGBM uses a leaf-wise tree growth that first expands the leaf with the best split score. This is the reason why the LightGBM is considered to be "light." The method for finding the best leaf split is a histogram-based algorithm. The feature values are segmented into bins, which are then utilized to construct the various feature histograms. These feature histograms yield the best split points for the leaves. Because the others are generated by subtracting its parent, the histograms only need to be updated for one leaf (the one with the least quantity of data) in each split (Ke et al., 2017).

According to the training trials for each algorithm, the test R^2 score becomes stabilized after the training sample rate reaches 70%. Thus, this research adopts the 70:30 training-test split ratio (301 samples for the test, 701 samples for training) in order to use as few training samples as possible to obtain as accurate predictions as possible.

3.5. Feature selection with PFI

The feature selection process is applied to determine the independent variables that contribute the most to the determination of resident prices and to eliminate the variables that do not contribute. In this way, the independent variables that are not useful to the predictive models are dropped, and the training of the models improves in terms of algorithmic speed, accuracy and processor usage (Kohavi & John, 1997). As indicated in the literature review, this study advocates the use of the PFI for feature selection (Breiman, 2001; Fisher, Rudin, & Dominici, 2019). This method seeks an answer to the following: How would the accuracy of predictions be affected if the analysts randomly shuffled any column of the test data while holding the dependent variable and other columns steady? Shuffling a single column should result in accuracy losses, and the model significantly loses accuracy if a column on which it relies for predictions is shuffled. Similarly, if any column that does not contribute significantly to the generation of the predictive model were shuffled, the subsequent predictions would not suffer nearly as much. The model accuracy in PFI analysis here is measured only with R^2 for test data.

To implement PFI procedure in alignment with the steps above, the ELI5 package of Python is used (ELI5, 2021). This package is designed to explain the predictions of ML models by assigning weights to each feature to represent PFI of the model. ELI5 is supported by Scikit-learn, LightGBM, and xgboost libraries. The ELI5 shuffles the feature variables many times and returns the average importance and standard deviation as output. Repeating the process with numerous shuffles yields the level of randomness (standard deviation) in PFI computation. The standard deviation measures how performance changed from one reshuffle to the next. When the average importance is close to zero, that feature variable contains little-to-no useful data. Thus, the feature variables with an average importance of less than 0.01 are to be dropped,[1] and the remaining feature variables are to be used for training the final regressor models.

3.6. Hyperparameter tuning

Each ML algorithm has its set of hyperparameters, and they need to be tuned before training the final regressors. The GridSearchCV procedure is used for hyperparameter tuning to determine the combination of parameters that yields the most accurate results. According to the results of the GridSearchCV procedure, the tuned hyperparameters are given as follows:

- RF: {*n_estimators*: 100}, {*max_features*: 5}, {*max_depth*: 5}, {*criterion*: Mean squared error}
- GB: {*learning_rate*: 0.01}, {*n_estimators*: 150}, {*max_features*: 0.5}, {*max_depth*: 6}, {*criterion*: Mean squared error}
- XGBoost: {*eta (learning_rate)*: 0.01}, {*max_depth*: 7}, {*alpha*: 0.01}, {*n_estimators*: 250}
- LightGBM: {*n_estimators*: 200}, {*max_depth*: 5}, {*learning_rate*: 0.01}, {*num_leaves*: 250}

3.7. Performance metrics

The performance of the predictors is detected with the metrics indicated in the report of *Standard on Ratio Studies* by IAAO (2013). These metrics are the Coefficient of Determination (R^2), the Mean Absolute Percentage Error (MAPE), the Root Mean Square Error (RMSE), the Coefficient of Dispersion (COD), and the Price-Related Differential (PRD).

$$R^2 = \left(\frac{\sum (y_i - \widehat{y})^2}{\sum (y_i - \overline{y})^2}\right) \tag{1}$$

$$MAPE = \frac{100\%}{n} \sum_{1}^{n} \frac{|y_i - \widehat{y}_i|}{|y_i|} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n} \Sigma_1^n (y_i - \widehat{y}_i)^2} \tag{3}$$

$$COD = \frac{\frac{100}{n}\sum_1^n |R_i - \widetilde{R}|}{\widetilde{R}} \tag{4}$$

$$PRD = \frac{\overline{R}}{Wt.\overline{R}} \tag{5}$$

In Eqs. (1)–(3), y denotes the actual value, $\widehat{y}$ denotes the appraisal value, and $\overline{y}$ denotes the mean of actual values. In Eq. (4), $R_i = \frac{A_i}{S_i}$ designates the ratio where A_i is the appraisal (predicted) value and S_i is the sale price of the sample. $\widetilde{R}$ corresponds to the median of R_i ratios in a data set with n

[1] There is no suggestion in the literature regarding this threshold value. In this study, the threshold value of 0.01 is chosen but it can be changed for other datasets and study areas.

numbers of samples. COD refers to the mean deviation from the median of a set of values represented as a percentage of the median. It is a widely used measure of appraisal uniformity. For residential units, the COD should not exceed 15. Last but not least, the PRD is an index statistic for measuring vertical equity between high-value and low-value properties. For single-family residential properties, it needs to be between 0.98 and 1.03. In Eq. (5), $\overline{R}$ refers to the mean of R_i ratios. $Wt.\overline{R}$ corresponds to the ratio of the sum of appraisal (predicted) values and the sum of sale prices $\left(\frac{\sum_1^n A_i}{\sum_1^n S_i}\right)$.

It is also important to assess the models in terms of their computational cost since a model demonstrating marginal improvements in accuracy at the cost of hours of training time is not particularly useful to practitioners or policy-makers. In ML, throughput - the number of samples processed in one unit of time (usually seconds) - can be used to assess the computational cost of different models. It is shown how best-performing models compare to traditional OLS in Table 3.

Finally, a pairwise Wilcoxon signed-rank test can be used to compare model outputs: since each had the same input features, significant differences in their predictions should be picked up by this test. This is a non-parametric test of signed ranks (Wilcoxon, 1945) that can be used to test the hypothesis that two sets of observations (the predictions from the ML models) were drawn from different probability distributions. In the Wilcoxon signed-rank test, the null hypothesis is that the collection of pairwise differences has a probability distribution centered at zero (Woolson, 2008).

3.8. *Local interpretation with SHAP*

Lastly, to interpret the output predicted by the algorithms, SHAP technique is applied (Shapley, 2016) using the *TreeExplainer* package (Lundberg & Lee, 2017). SHAP technique helps the analyst to understand the relevance of input features, as well as their reciprocal relationships (Iban & Sekertekin, 2022). SHAP is theoretically based on game theory and explains the output of ML models in accordance with the values of selected features. SHAP values, in other words, reflect the contribution of all selected features to the final appraisal (Kumar, Choudary, Bommineni, Tarun, & Anjali, 2020).

SHAP values are calculated for a specific feature by comparing the predicted values when the feature value is available against when it is hidden. Gathering model predictions for all conceivable subsets of features that contain and exclude the feature of interest is required for the precise calculation. As a result, the weight of each feature is computed. The resulting SHAP value illustrates how a model acts on the target value as well as how the independent feature values within the model are associated with one another (Wojtuch, Jankowski, & Podlewska, 2021; Yamaguchi, 2020).

4. Results and discussions

4.1. *Selected features*

In Table 2, the quantitative results from PFI analysis are presented. This table shows the selected features whose importance scores exceed the 0.01 threshold value. The numbers after the ± correspond to the amount of randomness in multiple permutations. With the help of PFI analysis, ML regressor models are no longer opaque boxes, and the intermediary steps and interactions between the features that have an impact on the predictions are crystal-clear. Hence, appraisal experts can investigate the features that go into the creation of a mass appraisal model, and the higher level of interpretability in a regressor model. In this study, *Gross Area* is the most important feature since PFI analysis shows that shuffling this feature decreases the predictive performance of tree-based regressor models by more than 20%. Furthermore, *Building Floors*, *Distance to Bazaars*, *Net Area*, *Number of Bathrooms* and *Number of*

Table 2
The permutation feature importance scores for each regressor.

RF			LightGBM			XGBoost			GB			OLS		
Features	Score	Randomness	Features	Score	Randomness	Features	Score	Randomness	Features	Score	Randomness	Features	Score	Randomness
Gross Area	0.2449	±0.0264	Gross Area	0.2009	±0.0125	Gross Area	0.3142	±0.0307	Gross Area	0.2009	±0.0125	Gross Area	0.1341	±0.0145
Number of Rooms	0.1022	±0.0123	Building Floors	0.0908	±0.0101	Distance to Bazaars	0.1247	±0.0071	Building Floors	0.0908	±0.0101	Building Floors	0.0866	±0.0242
Building Floors	0.0980	±0.0108	Distance to Bazaars	0.0589	±0.0067	Building Floors	0.1185	±0.0185	Distance to Bazaars	0.0589	±0.0067	Elevation	0.0839	±0.0174
Distance to Bazaars	0.0917	±0.0095	Number of Bathrooms	0.0482	±0.0090	Number of Bathrooms	0.0547	±0.0097	Number of Bathrooms	0.0482	±0.0090	Distance to Rail Station	0.0715	±0.0091
Number of Bathrooms	0.0769	±0.0098	Number of Rooms	0.0477	±0.0127	Net Area	0.0438	±0.0038	Number of Rooms	0.0477	±0.0127	Distance to Military Zones	0.0571	±0.0066
Net Area	0.0335	±0.0029	Building Age	0.0334	±0.0052	Building Age	0.0428	±0.0095	Building Age	0.0334	±0.0052	Building Age	0.0554	±0.0050
Building Age	0.0277	±0.0057	Net Area	0.0318	±0.0052	Number of Rooms	0.0355	±0.0059	Net Area	0.0318	±0.0052	Net Area	0.0547	±0.0062
Elevation	0.0248	±0.0043	Distance to Coach Station	0.0247	±0.0051	Elevation	0.0321	±0.0041	Distance to Coach Station	0.0247	±0.0051	Distance to Bazaars	0.0513	±0.0190
Distance to Coastline	0.0167	±0.0044	Distance to Industrial Areas	0.0183	±0.0041	Distance to Coach Station	0.0291	±0.0027	Distance to Industrial Areas	0.0183	±0.0041	Distance to Coastline	0.0323	±0.0106
Distance to Coach Station	0.0122	±0.0012	Elevation	0.0139	±0.0034	Heating System	0.0278	±0.0070	Elevation	0.0139	±0.0034	Distance to Port	0.0254	±0.0080
Floor Number	0.0115	±0.0003	Heating System	0.0129	±0.0022	Floor Number	0.0231	±0.0023	Heating System	0.0129	±0.0022	Heating System	0.0195	±0.0077
									Floor Number	0.0118	±0.0027	Number of Bathrooms	0.0119	±0.0023
												Distance to Parks	0.0105	±0.0037

Rooms have an effect of more than 1 percent in all regressor models. Some features selected as important in the OLS model were not found to be important for tree-based models. Tree-based models attribute importance to almost the same feature variables.

4.2. Predictive performance of the regressors

Using selected features and tuned hyperparameters, the GB performed the highest R^2 score as illustrated in Table 3. However, all tree-based models have an acceptable R^2 score of more than 0.9. The tree-based models have MAPE values of between 0.06 and 0.14. The XGBoost shows the lowest MAPE and RMSE values. In terms of COD and PRD, XGBoost and LightGBM show better performances than the other models. Most importantly, COD and PRD values satisfy the limits of IAAO standards. On the other hand, the OLS regressor did not perform as satisfactory as the tree-based models in terms of R^2, MAPE, and RMSE metrics. COD and PRD values obtained from the OLS are not at an acceptable level that international standards address. In this context, it is possible to say that the tree-based models show better prediction accuracy performance than the classical OLS approach.

To compare with the studies conducted by other scholars, MAPE and R^2 results are cross-checked since they are the most common performance metrics in the existing body of literature. The RF model in this study outperforms some of the RF models generated by other scholars (Antipov & Pokryshevskaya, 2012; Alfaro-Navarro et al., 2020; Yilmazer & Kocaman, 2020; Bilgilioğlu & Yılmaz, 2021; Ho et al., 2021; Tchuente & Nyawa, 2022)). However, the RMSE and MAPE of the RF models generated by Aydinoglu et al. (2021) and Čeh et al. (2018) are slightly better than the performance of the RF model generated in this study. Moreover, the GB model from this study outperforms the GB models generated by Tchuente & Nyawa (2022) and Antipov and Pokryshevskaya (2012). Since the XGBoost and LightGBM regressors have not been implemented so frequently in the MAS studies, they cannot be compared. However, they seem to be reliable methods with a prediction score close to that of the RF learner. These outcomes support the idea that tree-based models need to take more attention to the MAS studies.

A Wilcoxon signed-rank test is used to see which regressors produce statistically significant predictions with a 95% confidence level (Wilcoxon, 1945). The p-values less than 0.05 indicate statistically significant evidence to reject the null hypothesis, and thus predictions generated by two regressors on the same real estate samples are statistically significant. If the p-values are greater than 0.05, the null hypothesis is cannot be rejected, and the predictions generated by two regressors are not statistically significant. According to the p-values derived from the Wilcoxon signed-rank test, the RF and XGBoost regressors generate statistically significant predictions. The p-value between other tree-based pairs is more than 0.05, and they are not statistically significant. The OLS makes a p-value less than 0.00001 once it is tested with other algorithms.

4.3. Local interpretation using SHAP values

For the local interpretation of the MAS model, the XGBoost regressor is chosen due to its slightly better performance. The bee swarm plot in Fig. 2a shows the range of SHAP values for each selected feature, arranged by significance. All the little dots on the plot represent a single sample. The horizontal axis represents SHAP value, while the color of the point shows us if that sample has a higher or a lower value when compared to other samples. The horizontal positioning of the points indicates whether they have a positive or negative impact on the prediction. Fig. 2b shows a simplified version of the bee swarm plot. It shows the correlation between the feature variables and absolute SHAP values. Green bars denote the features having a positive impact on the predictions, while the red ones refer to negative impact.

The plots show that gross area, number of bathrooms, number of rooms, and building floors are the most contributing features. This outcome is not surprising since the main drivers of the transaction prices are generally physical-based (Jafari & Akhavian, 2019; Chun Lin & Mohan, 2011). Similar to the findings of Stamou, Mimis, and Rovolis (2017) for Athens, the residents of Mersin seem to prefer higher buildings and higher floor numbers, supporting the hypothesis of vertical differentiation. Among the location-based features, distance to bazaars and distance to coach station are important for the Yenişehir market. Location-based features should be interpreted correctly here. Longer (high value) distances to bazaars and shorter (low value) distances to coach stations increase the possibility of obtaining a higher residential price. This does not mean that the bazaar is a pushing factor and the coach station is an attraction. This situation shows that buying and selling transactions may be realized at higher prices in areas that do not have a bazaar yet or are close to the coach station. In the literature, there are many studies that show the positive contribution of factors such as proximity to public services, education and health facilities, roads, green areas, public transportation and the seaside on housing prices (Daams et al., 2019; Giannico et al., 2021; Hui et al., 2007; Kang, Zhang, Gao, et al., 2021; Kang, Zhang, Peng, et al., 2021; Su et al., 2021; Wang et al., 2017; Xiao et al., 2019; Yuan et al., 2020). The fact that these factors are not important in the model developed for the Mersin housing market does not mean that these factors are not important for the individuals. In Mersin city, as a result of the homogeneous distribution of these facilities, the distance to these factors is not a distinguishing feature of the prices. Thanks to this homogeneous distribution, those who will buy a house do not doubt whether the house they will buy is close to the school. However, if a newly developing neighborhood does not have access to education or health services for a long time, house prices in that neighborhood are expected to be adversely affected. In short, the contribution of the factors to the model or their importance in the model explains not the degree of importance of those factors for the individuals, but whether they have a distinctive role in determining house prices.

The bee swarm plot gives an idea about the factors affecting the house price, but in a global manner. However, with this graph, it is not possible to obtain information about the factors that affect the prediction of the appraisal value of a single sample. Therefore, it is necessary to draw a force plot showing the magnitude of SHAP values for each sample (Chen et al., 2020). In Fig. 3, force plots can be seen for two randomly selected samples. The force plot is another way to see the effect of each feature has on the prediction, for a given sample. In these plots, the positive SHAP values are displayed on the left side and the negative on the right side, as if competing against each other. The highlighted value in the middle is the prediction for that sample. SHAP

Table 3
Performance results of each regressor model.

Regressors	R^2	MAPE	RMSE (Turkish Liras)	COD	PRD	Training Throughput (Samples per second)	Prediction Throughput (Samples per second)
RF	0.8957	0.1056	158,671	10.3119	1.029	191.44	6272.09
GB	**0.9052**	0.1410	193,902	14.2383	**0.9942**	702.22	40133.69
XGBoost	0.8733	**0.0904**	**146,837**	**8.9802**	1.0002	527.80	33771.65
LightGBM	0.8961	0.0985	154,416	9.8851	**0.9942**	1497.99	45444.21
OLS	0.7510	0.2141	270,091	17.4039	1.0344	**3525.41**	**50170.30**

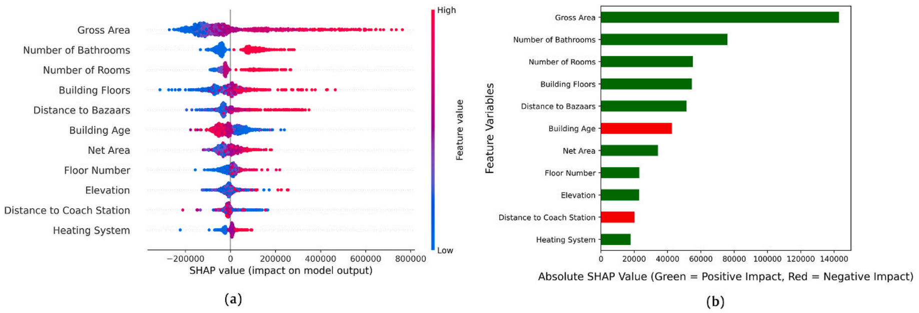

Fig. 2. (a) Bee swarm plot of SHAP values generated by the XGBoost, (b) a simplified version of bee swarm plot.

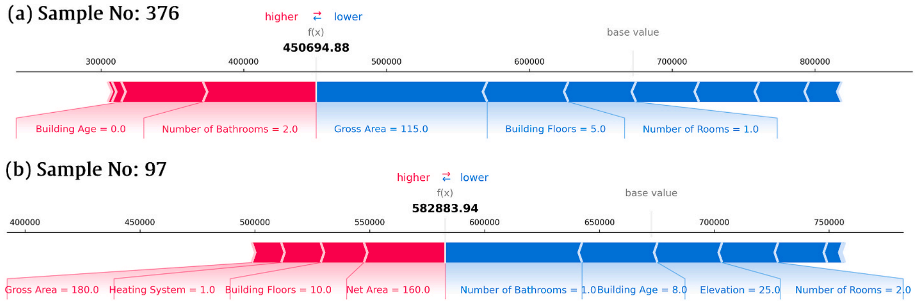

Fig. 3. Shap force plots for two randomly selected samples (a) sample 376, (b) sample 97.

values are attribution methods so that the prediction of a single sample is described as the sum of feature effects. For sample No: 376 (Fig. 3a), the gross area of 115 square meters generated a negative SHAP value, which negatively affected the appraisal value. However, in the other sample (No:97 in Fig. 3b), the gross area of 180 square meters created a SHAP value that positively affected the appraisal value. Moreover, in the first sample, having two bathrooms in a one-room residence created a SHAP value that contributed positively to the appraisal value. However, in the second sample, in a larger two-bedroom residence, having a single bathroom produced a negative SHAP value that lowered the appraisal value. On the other hand, it is possible to see the difference between the appraisal value and the base value (the original input) on force plots. Once the difference is very large, the feature values might have bias. The analyst can detect these significant differences (underestimates or overestimates) and check SHAP values of features that may cause such differences.

These are the advantages of locally explainable ML methods that this study tries to highlight. Although it is not identified as an important factor in this study, there are studies that have determined that the presence of an elevator in a building has an absolute effect on the appraisal value. Globally explainable ML methods can give a higher importance score for the existence of the elevator. However, global methods cannot determine whether the presence of elevators affects the appraisal value in one-storey or two-storey buildings in the database. In other words, the elevator can be an important criterion throughout the city. However, in very low-rise buildings, the absence of an elevator may not negatively affect the appraisal value. To see this, it is necessary to monitor the samples locally, as in the force plot example. Hence, the MAS using locally XAI methods will also help to determine the location-specific coefficients used for individual appraisals.

MAS models cannot survive with static data sets. Therefore, in order to maintain the global and local explanation of the MAS models, data sets should be kept up-to-date and instant preferences should be monitored. The dataset used in this study depicts the market preferences in Mersin for the first six months of 2021. However, in the following years, migration, natural disasters, epidemics, macro-economic conditions and political changes may cause different factors to affect the real estate values throughout the city or cause the factors that are important in their current form to become unimportant. Lockdowns have transformed the way individuals and communities live, engage, and work since the COVID-19 pandemic emerged. It also teaches us the importance of making the built environment resilient (Wu, Zheng, & Li, 2022). Unprecedented developments and world-changing events can quickly change preferences in housing markets. In order not to be caught unprepared for such situations, an MAS mechanism working with fast and high accuracy algorithms, and based on explainable methods help policy- and decision-makers to respond quickly to a rapidly changing environment.

5. Conclusions

This study contributes toward a better understanding of Explainable Artificial Intelligence (XAI) in Mass Appraisal Systems (MAS) of properties specifically, and in urban studies more generally. Moreover, the content of this study may stimulate the debate on the use of tree-based Machine Learning (ML) algorithms for MAS and tries to set up a regression scheme using RF, XGBoost, GB, and LightGBM algorithms. The experimental results derived from this study show that tree-based algorithms can be indispensable tools in MAS studies and can capture the delicate and complicated interactions between features. The predictive performance of tree-based algorithms outperformed that of the classical OLS method, and the XGBoost regressor seems to be slightly better than competing tree-based methods.

The main research question of this study is to assess the explainability of the ML-based MAS studies. The MAS databases may contain dozens of independent variables, and selecting the most important ones for a specific area is a challenging task for appraisers. The analyst may be inclined to use surveys to define the trends and behaviors in the market, but this can be an expensive and time-consuming approach. Thanks to ML algorithms, it is now possible to appraise multiple properties using a set of training data. However, stand-alone ML regressors are known to be black-box models, which do not allow the analyst to interpret the contribution of each input variable to the price predictions. To eliminate this problem, this study suggests the use of Permutation Feature

Importance (PFI) for feature selection to detect the most important factors for pricing. This study finds PFI is a feasible solution to feature selection from complex, correlated data sets.

However, the methodology of this study does not find global XAI methods sufficient. In order to answer the research question comprehensively, each real estate sample should also be explained locally. In this respect, SHAP method has been used and the effect of value determinants on the price prediction of each sample has been determined. For two randomly selected samples, it has been observed that the value variables have different SHAP values. In other words, all the predictions in the generated MAS model take into account the independent variables differently.

The contributions of this study to the literature can be summarized under three headings.

- This study claims that tree-based regressors give stronger results in terms of predictive performance. Therefore, it encourages analysts to use tree-based algorithms, especially gradient-based ones, when developing their models.
- It reports that PFI method is a faster and more robust option, and recommends that this method be used more widely for feature selection.
- With SHAP analysis, a local explanation was provided for each prediction. In this way, analysts working on large data sets can monitor the value determinants for each of the real estate properties they are interested in. With an updated database, they can follow the value determinants of the market on a sample-wise scale and compare them with social, demographic, and economic indicators.

Of course, there are some limitations to be addressed in future studies:

- This study carried out the experiments on a data set collected in only one district. Testing of the model in other cities could not be achieved due to a lack of data.
- The reliability of the collected samples is not very high. Therefore, outlier analysis could not be performed. In order for the model to make more robust predictions, it is necessary to work with verified and reliable value data.
- There are only physical and spatial variables in the data set. Including socio-economic and demographic data sets that affect the value of residences in such an analysis will help create a more realistic MAS.

Data and code availability

The data sets and codes are available on reasonable request.

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.

Acknowledgments

We thank the editor and three anonymous reviewers for their constructive contributions and the time they have spent on this article.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alfaro-Navarro, J.-L., Cano, E. L., Alfaro-Cortés, E., García, N., Gámez, M., & Larraz, B. (2020). A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems. *Complexity*, 1–12. https://doi.org/10.1155/2020/5287263, 2020.

Alpaydin, E. (2020). *Introduction to machine learning.* MIT Press.

Amarasinghe Arachchige, J., Quach, S., Roca, E., Liu, B., Liew, A. W., & Earl, G. (2022). Understanding high-involvement product purchase through an innovative machine learning approach: A case of housing type choice. *Journal of Consumer Behaviour*, 1–18. https://doi.org/10.1002/cb.2055. In press.

Amparore, E., Perotti, A., & Bajardi, P. (2021). To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods. *PeerJ Computer Science, 7*, e479. https://doi.org/10.7717/peerj-cs.479

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications, 39*(2), 1772–1778. https://doi.org/10.1016/j.eswa.2011.08.077

Aydinoglu, A. C., Bovkir, R., & Colkesen, I. (2021). Implementing a mass valuation application on interoperable land valuation data model designed as an extension of the national GDI. *Survey Review, 53*(379), 349–365. https://doi.org/10.1080/00396265.2020.1771967

Bartke, S., & Schwarze, R. (2021). The economic role and emergence of professional valuers in real estate markets. *Land, 10*(7), 683. https://doi.org/10.3390/land10070683

Bilgilioğlu, S. S., & Yılmaz, H. M. (2021). Comparison of different machine learning models for mass appraisal of real estate. *Survey Review*, 1–12. https://doi.org/10.1080/00396265.2021.1996799

Bin, J., Gardiner, B., Li, E., & Liu, Z. (2020). Multi-source urban data fusion for property value assessment: A case study in philadelphia. *Neurocomputing, 404*, 70–83. https://doi.org/10.1016/j.neucom.2020.05.013

Bishop, C. (2006). *Pattern recognition and machine learning.* Springer-Verlag.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Bunyan Unel, F., & Yalpir, S. (2019). Valuations of building plots using the AHP method. *International Journal of Strategic Property Management, 23*(3), 197–212. https://doi.org/10.3846/ijspm.2019.7952

Çağdaş, V. (2013). An application domain extension to CityGML for immovable property taxation: A Turkish case study. *Int. J. Appl. Earth Obs. Geoinformation, 21*, 545–555. https://doi.org/10.1016/j.jag.2012.07.013

Cascarino, G., Moscatelli, M., & Parlapiano, F. (2022). Explainable artificial intelligence: Interpreting default forecasting models based on machine learning. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.4090707

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information, 7*(5), 168. https://doi.org/10.3390/ijgi7050168

Chaturvedi, V., & de Vries, W. T. (2021). Machine learning algorithms for urban land use planning: A review. *Urban Science, 5*(3), 68. https://doi.org/10.3390/urbansci5030068

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '16* (pp. 785–794). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.

Chen, J.-H., Ong, C. F., Zheng, L., & Hsu, S.-C. (2017). Forecasting spatial dynamics of the housing market using Support Vector Machine. *International Journal of Strategic Property Management, 21*(3), 273–283. https://doi.org/10.3846/1648715X.2016.1259190

Chen, L., Yao, X., Liu, Y., Zhu, Y., Chen, W., Zhao, X., et al. (2020). Measuring impacts of urban environmental elements on housing prices based on multisource data—a case study of shanghai, China. *ISPRS International Journal of Geo-Information, 9*(2), 106. https://doi.org/10.3390/ijgi9020106

Chun Lin, C., & Mohan, S. B. (2011). Effectiveness comparison of the residential property mass appraisal methodologies in the USA. *International Journal of Housing Markets and Analysis, 4*(3), 224–243. https://doi.org/10.1108/17538271111153013

Colwell, P. F., & Dilmore, G. (1999). Who was first? An examination of an early hedonic study. *Land Economics, 75*(4), 620. https://doi.org/10.2307/3147070

Connellan, O., & James, H. (1998). Estimated realisation price (ERP) by neural networks: Forecasting commercial property values. *Journal of Property Valuation and Investment, 16*, 71–86. https://doi.org/10.1108/14635789810205137

Daams, M. N., Sijtsma, F. J., & Veneri, P. (2019). Mixed monetary and non-monetary valuation of attractive urban green space: A case study using amsterdam house prices. *Ecological Economics, 166*, Article 106430. https://doi.org/10.1016/J.ECOLECON.2019.106430

Das, A., & Rad, P. (2020). *Opportunities and challenges in explainable artificial intelligence (XAI): A survey.* https://doi.org/10.48550/arxiv.2006.11371

Delgado-Panadero, Á., Hernández-Lorca, B., García-Ordás, M. T., & Benítez-Andrades, J. A. (2022). Implementing local-explainability in gradient boosting trees: Feature contribution. *Information Sciences, 589*, 199–212. https://doi.org/10.1016/J.INS.2021.12.111

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1–15). Berlin, Heidelberg: Springer Berlin Heidelberg.

Doumard, E., Aligon, J., Escriva, E., Excoffier, J.-B., Monsarrat, P., & Soulé-Dupuy, C. (2022). A comparative study of additive local explanation methods based on feature influences. *24th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data*, 31–40. *DOLAP 2022), 3130*(paper 4) https://hal.archives-ouvertes.fr/hal-03687554.

Doumpos, M., Papastamos, D., Andritsos, D., & Zopounidis, C. (2021). Developing automated valuation models for estimating property values: A comparison of global and locally weighted approaches. *Annals of Operations Research, 306*(1–2), 415–433. https://doi.org/10.1007/s10479-020-03556-1

ELI5. (2021). Teamhg-Memex/eli5, 07.04.22 https://github.com/TeamHG-Memex/eli5.

Fedorov, N., & Petrichenko, Y. (2020). Gradient boosting–based machine learning methods in real estate market forecasting. In *Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)* (pp. 203–208). https://doi.org/10.2991/aisr.k.201029.039

Fields, D., & Rogers, D. (2021). Towards a critical housing studies research agenda on platform real estate. *Housing, Theory and Society, 38*(1), 72–94. https://doi.org/10.1080/14036096.2019.1670724

Filippakopoulou, M., & Potsiou, C. (2014). Research on residential property taxation and its impact on the real estate market in Greece. *Survey Review, 46*(338), 333–341. https://doi.org/10.1179/1752270614Y.0000000113

Fisher, A. J., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res. JMLR, 20*.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Ghalebikesabi, S., Ter-Minassian, L., DiazOrdaz, K., & Holmes, C. C. (2021). On locality of local explanation models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 18395–18407). Curran Associates, Inc.. https://proceedings.neurips.cc/paper/2021/file/995665640dc319973d3173a74a03860c-Paper.pdf

Giannico, V., Spano, G., Elia, M., D'Este, M., Sanesi, G., & Lafortezza, R. (2021). Green spaces, quality of life, and citizen perception in European cities. *Environmental Research, 196*, Article 110922. https://doi.org/10.1016/J.ENVRES.2021.110922

Glumac, B., & des Rosiers, F. (2021). Practice briefing – automated valuation models (AVMs): Their role, their advantages and their limitations. *Journal of Property Investment & Finance, 39*(5), 481–491. https://doi.org/10.1108/JPIF-07-2020-0086

Gnat, S. (2021). Property mass valuation on small markets. *Land, 10*(4), 388. https://doi.org/10.3390/land10040388

González, S., García, S., Ser, J. D., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion, 64*, 205–237. https://doi.org/10.1016/j.inffus.2020.07.007

Grekousis, G. (2019). Artificial neural networks and deep learning in urban geography: A systematic review and meta-analysis. *Computers, Environment and Urban Systems, 74*, 244–256. https://doi.org/10.1016/j.compenvurbsys.2018.10.008

Grover, R. (2016). Mass valuations. *Journal of Property Investment & Finance, 34*(2), 191–204. https://doi.org/10.1108/JPIF-01-2016-0001

Hamilton, S. E., & Morgan, A. (2010). Integrating lidar, GIS and hedonic price modeling to measure amenity values in urban beach residential property markets. *Computers, Environment and Urban Systems, 34*(2), 133–141. https://doi.org/10.1016/j.compenvurbsys.2009.10.007

Hass, G. C. (1922). *Sale prices as a basis for farm land appraisal (No. Technical bulletin 9)*.

Hefferan, M. J., & Boyd, T. (2010). Property taxation and mass appraisal valuations in Australia – adapting to a new environment. *Property Management, 28*(3), 149–162. https://doi.org/10.1108/02637471011051291

Hei-Ling Lam, C., & Chi-Man Hui, E. (2018). How does investor sentiment predict the future real estate returns of residential property in Hong Kong? *Habitat International, 75*, 1–11. https://doi.org/10.1016/j.habitatint.2018.02.009

Hong, J., Choi, H., & Kim, W. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management, 24*(3), 140–152. https://doi.org/10.3846/ijspm.2020.11544

Ho, W. K. O., Tang, B.-S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research, 38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Huang, N., Lu, G., & Xu, D. (2016). A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. *Energies, 9*(10), 767. https://doi.org/10.3390/en9100767

Hui, E. C. M., Chau, C. K., Pun, L., & Law, M. Y. (2007). Measuring the neighboring and environmental effects on residential property value: Using spatial weighting matrix. *Building and Environment, 42*(6), 2333–2343. https://doi.org/10.1016/j.buildenv.2006.05.004

IAAO. (2013). *Standard on ratio studies*, 07.04.22 https://www.iaao.org/media/standards/Standard_on_Ratio_Studies.pdf.

Iban, M. C., & Sekertekin, A. (2022). Machine learning based wildfire susceptibility mapping using remotely sensed fire data and GIS: A case study of Adana and Mersin provinces, Turkey. *Ecological Informatics, 69*, Article 101647. https://doi.org/10.1016/j.ecoinf.2022.101647

Islam, S. R., Eberle, W., Ghafoor, S. K., & Ahmed, M. (2021). *Explainable artificial intelligence approaches: A survey*. https://doi.org/10.48550/arxiv.2101.09429

Jafari, A., & Akhavian, R. (2019). Driving forces for the US residential housing price: A predictive analysis. *Built Environment Project and Asset Management, 9*(4), 515–529. https://doi.org/10.1108/BEPAM-07-2018-0100

Jia, J., Zhang, X., Huang, C., & Luan, H. (2022). Multiscale analysis of human social sensing of urban appearance and its effects on house price appreciation in Wuhan, China. *Sustainable Cities and Society, 81*, Article 103844. https://doi.org/10.1016/j.scs.2022.103844

Kaczmarek, I., Iwaniak, A., Świetlicka, A., Piwowarczyk, M., & Nadolny, A. (2022). A machine learning approach for integration of spatial development plans based on natural language processing. *Sustainable Cities and Society, 76*, Article 103479. https://doi.org/10.1016/j.scs.2021.103479

Kang, Y., Zhang, F., Gao, S., Peng, W., & Ratti, C. (2021). Human settlement value assessment from a place perspective: Considering human dynamics and perceptions in house price modeling. *Cities, 118*, Article 103333. https://doi.org/10.1016/J.CITIES.2021.103333

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., et al. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy, 111*, Article 104919. https://doi.org/10.1016/J.LANDUSEPOL.2020.104919

Kathmann, R. M. (1993). Neural networks for the mass appraisal of real estate. *Computers, Environment and Urban Systems, 17*(4), 373–384. https://doi.org/10.1016/0198-9715(93)90034-3

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st international conference on neural information processing systems, NIPS'17* (pp. 3149–3157). Red Hook, NY, USA: Curran Associates Inc.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *Journal of Portfolio Management, 43*(6), 202–211. https://doi.org/10.3905/jpm.2017.43.6.202

Konstantinov, A.v., & Utkin, L.v. (2021). Interpretable machine learning with an ensemble of gradient boosting machines. *Knowledge-Based Systems, 222*. https://doi.org/10.1016/J.KNOSYS.2021.106993

Kumar, C. S., Choudary, M. N. S., Bommineni, V. B., Tarun, G., & Anjali, T. (2020). Dimensionality reduction based on shap analysis: A simple and trustworthy approach. In *2020 international conference on communication and signal processing (ICCSP)* (pp. 558–560). IEEE. https://doi.org/10.1109/ICCSP48568.2020.9182109.

Leao, S. Z., van den Nouwelant, R., Shi, V., Han, H., Praharaj, S., & Pettit, C. J. (2021). A rapid analytics tool to map the effect of rezoning on property values. *Computers, Environment and Urban Systems, 86*, Article 101572. https://doi.org/10.1016/j.compenvurbsys.2020.101572

Lenk, M. M., Worzala, E. M., & Silva, A. (1997). High-tech valuation: Should artificial neural networks bypass the human valuer? *Journal of Property Valuation and Investment, 15*, 8–26. https://doi.org/10.1108/14635789710163775

Li, X., Hui, E. C. M., & Shen, J. (2020). The consequences of Chinese outward real estate investment: Evidence from Hong Kong land market. *Habitat International, 98*, Article 102151. https://doi.org/10.1016/j.habitatint.2020.102151

Ling, Z., & Hui, E. C. M. (2013). Structural change in housing submarkets in burgeoning real estate market: A case of hangzhou, China. *Habitat International, 39*, 214–223. https://doi.org/10.1016/j.habitatint.2012.12.006

Lo, D., Chau, K. W., Wong, S. K., McCord, M., & Haran, M. (2022). Factors affecting spatial autocorrelation in residential property prices. *Land, 11*(6), 931. https://doi.org/10.3390/land11060931

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

McCluskey, W., Deddis, W., Mannis, A., McBurney, D., & Borst, R. (1997). Interactive application of computer assisted mass appraisal and geographic information systems. *Journal of Property Valuation and Investment, 15*(5), 448–465. https://doi.org/10.1108/14635789710189227

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: A comparison of modern approaches. *Journal of Property Research, 30*(4), 239–265. https://doi.org/10.1080/09599916.2013.781204

McCluskey, W., Zulkarnain Daud, D., & Kamarudin, N. (2014). Boosted regression trees: An application for the mass appraisal of residential property in Malaysia. *Journal of Financial Management of Property and Construction, 19*(2), 152–167. https://doi.org/10.1108/JFMPC-06-2013-0022

Ming, Y. S., & Hin, H. K. (2006). Planned urban industrialization and its effect on urban industrial real estate valuation: The Singapore experience. *Habitat International, 30*(3), 509–539. https://doi.org/10.1016/j.habitatint.2004.12.006

Pai, P.-F., & Wang, W.-C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences, 10*(17), 5832. https://doi.org/10.3390/app10175832

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Resch, B., & Szell, M. (2019). Human-centric data science for urban studies. *ISPRS International Journal of Geo-Information, 8*(12), 584. https://doi.org/10.3390/ijgi8120584

Samek, W. (2020). Learning with explainable trees. *Nature Machine Intelligence, 2*(1), 16–17. https://doi.org/10.1038/s42256-019-0142-0

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*(2), 2843–2852. https://doi.org/10.1016/j.eswa.2008.01.044

Shapley, L. (2016). 17. A value for n-person games. In H. Kuhn, & A. Tucker (Eds.), *Contributions to the theory of games (AM-28)* (Vol. II, pp. 307–318). Princeton: Princeton University Press. https://doi.org/10.1515/9781400881970-018.

Sisman, S., & Aydinoglu, A. C. (2022). Improving performance of mass real estate valuation through application of the dataset optimization and Spatially Constrained Multivariate Clustering Analysis. *Land Use Policy, 119*, Article 106167. https://doi.org/10.1016/j.landusepol.2022.106167

Stamou, M., Mimis, A., & Rovolis, A. (2017). House price determinants in Athens: A spatial econometric approach. *Journal of Property Research, 34*(4), 269–284. https://doi.org/10.1080/09599916.2017.1400575

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Su, S., He, S., Sun, C., Zhang, H., Hu, L., & Kang, M. (2021). Do landscape amenities impact private housing rental prices? A hierarchical hedonic modeling approach based on semantic and sentimental analysis of online housing advertisements across five Chinese megacities. *Urban Forestry and Urban Greening, 58*, Article 126968. https://doi.org/10.1016/J.UFUG.2020.126968

Suparman, Y., Folmer, H., & Oud, J. H. L. (2014). Hedonic price models with omitted variables and measurement errors: A constrained autoregression–structural equation modeling approach with application to urban Indonesia. *Journal of Geographical Systems, 16*(1), 49–70. https://doi.org/10.1007/s10109-013-0186-3

Taecharungroj, V. (2021). Google Maps amenities and condominium prices: Investigating the effects and relationships using machine learning. *Habitat International, 118*, Article 102463. https://doi.org/10.1016/j.habitatint.2021.102463

Tajani, F., Morano, P., & Ntalianis, K. (2018). Automated valuation models for real estate portfolios. *Journal of Property Investment & Finance, 36*(4), 324–347. https://doi.org/10.1108/JPIF-10-2017-0067

Tchuente, D., & Nyawa, S. (2022). Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research, 308*(1–2), 571–608. https://doi.org/10.1007/s10479-021-03932-5

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). *Evaluating XAI: A comparison of rule-based and example-based explanations* (Vol. 291). Artificial Intelligence. https://doi.org/10.1016/j.artint.2020.103404

Wang, X. R., Hui, E. C. M., & Sun, J. X. (2017). Population migration, urbanization and housing prices: Evidence from the cities in China. *Habitat International, 66*, 49–56. https://doi.org/10.1016/J.HABITATINT.2017.05.010

Wang, D., & Li, V. J. (2019). Mass appraisal models of real estate in the 21st century: A systematic literature review. *Sustainability, 11*(24), 7006. https://doi.org/10.3390/su11247006

Watson, D. (2019). The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines, 29*(3), 417–440. https://doi.org/10.1007/s11023-019-09506-6

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometric Bulletin, 1*(6), 80. https://doi.org/10.2307/3001968

Wojtuch, A., Jankowski, R., & Podlewska, S. (2021). How can SHAP values help to shape metabolic stability of chemical compounds? *Journal of Cheminformatics, 13*, 74. https://doi.org/10.1186/s13321-021-00542-y

Woolson, R. F. (2008). Wilcoxon signed-rank test. In *Wiley encyclopedia of clinical trials*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780471462422.eoct979.

Wu, Q., Zheng, Z., & Li, W. (2022). Can housing assets affect the Chinese residents' willingness to pay for green housing? *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.782035

Xiao, Y., Hui, E. C. M., & Wen, H. (2019). Effects of floor level and landscape proximity on housing price: A hedonic analysis in hangzhou, China. *Habitat International, 87*, 11–26. https://doi.org/10.1016/J.HABITATINT.2019.03.008

Xu, X., Qiu, W., Li, W., Liu, X., Zhang, Z., Li, X., et al. (2022). Associations between street-view perceptions and housing prices: Subjective vs. Objective measures using computer vision and machine learning techniques. *Remote Sensing, 14*(4), 891. https://doi.org/10.3390/rs14040891

Yalpir, S., Sisman, S., Akar, A. U., & Unel, F. B. (2021). Feature selection applications and model validation for mass real estate valuation systems. *Land Use Policy, 108*, Article 105539. https://doi.org/10.1016/j.landusepol.2021.105539

Yamaguchi, K. (2020). Intrinsic meaning of shapley values in regression. In *2020 11th international conference on awareness science and technology (ICAST)* (pp. 1–6). https://doi.org/10.1109/iCAST51195.2020.9319492

Yang, J. P., & Bai, Q. (2013). Research of real estate appraisal based on GIS technology. *Advanced Materials Research, 859*, 562–565. https://doi.org/10.4028/www.scientific.net/AMR.859.562.

Yilmazer, S., & Kocaman, S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. *Land Use Policy, 99*, Article 104889. https://doi.org/10.1016/j.landusepol.2020.104889

Yuan, F., Wei, Y. D., & Wu, J. (2020). Amenity effects of urban facilities on housing prices in China: Accessibility, scarcity, and urban spaces. *Cities, 96*, Article 102433. https://doi.org/10.1016/J.CITIES.2019.102433