PRINT ISSN 1119-8362
Electronic ISSN 2659-1499

Full-text Available Online at
https://www.ajol.info/index.php/jasem
https://www.bioline.org.br/ja

J. Appl. Sci. Environ. Manage.
Vol. 29 (1) 43-48 Jan. 2025

Evaluation of Price Prediction of Houses in a Real Estate via Machine Learning

[1]*IBRAHIM, AA; [2]AYILARA-ADEWALE, OA; [2]ALABI, AA; [4]OLUSESI, DA

*[1]*Department of Mathematical Sciences, [4]Department of Mathematical Sciences, Oduduwa University, Ipetumodu, Nigeria*
[2]Department of Information Technology, [3]Department of Information Science, Osun State University, Osogbo, Nigeria

**Corresponding Author Email: adebisiibrahim97@gmail.com*
**ORCID: https://orcid.org/0000-0002-8797-1369*
**Tel: +2347034845265*

Co-Authors Emails: ayilara_adewale@uniosun.edu.ng; akeem.alabi@uniosun.edu.ng; daniel.olusesi20@gmail.com

ABSTRACT: Traditional (manual) methods of determining real estate house prices are in some cases prone to mistakes which may be due to distractions, lack of attentiveness or vulnerability to real estate agent fraud. This work focuses on evaluation house price prediction in real estate using more recent methods. House pricing using such methods as House Pricing Index and Random Forest Machine Learning Technique has been discussed, a new approach is proposed as a model utilizing the Extra Tree regression because it introduces an additional level of randomness in the tree-building process. Kaggle Boston housing dataset with 506 entries and 14 features was employed to train and test the developed model whose efficiency was then determined via mean absolute error and mean squared error. Additionally, a comparison was made between a random forest regression model and the proposed prediction model which revealed that the new prediction model yielded better performance than the random forest regression.

DOI: https://dx.doi.org/10.4314/jasem.v29i1.6

Cite this Article as: IBRAHIM, AA; AYILARA-ADEWALE, OA; ALABI, AA; OLUSESI, D. A. (2025). Evaluation of Price Prediction of Houses in a Real Estate via Machine Learning. *J. Appl. Sci. Environ. Manage.* 29 (1) 43-48

Dates: Received: 22 October 2024; Revised: 20 November 2024; Accepted: 08 December 2024; Published: 31 December 2024

Keywords: Prediction system; used car; extra tree regression; random forest regression; machine learning

The real estate, a property consisting of land as well as the buildings erected on it with its natural resources (Cheng, 2023), as a market constitutes a significant segment of the global economy, and housing stands as one of the most valuable assets for individuals and families. The demand for accommodation and industrial buildings rises with population increase. House prices are generally affected by several factors such as the location and size of the facility, access to basic infrastructure and amenities which include good roads, electricity, hospitals, fire agencies and markets, security, as well as other notable socio-economic indicators or determinants such as economic conditions and demand (Thamarai *et al*, 2020, Akhil, *et al*, 2023). Accurate prediction of house prices holds paramount importance as it provides crucial insights for buyers, sellers, real estate agents, and financial institutions, empowering them to make well-informed decisions regarding property investments **(**Zemlyanskiy, 2021**).** Real estate encompasses a broad spectrum of properties, including residential, commercial, and industrial spaces. It constitutes an integral sector within the economy generating substantial revenue and employment opportunities (Kuvalekar *et al*, 2020). A distinct feature of the real estate market includes its dynamic nature where the values of the properties can fluctuate as a result of internal and external factors. Given the substantial financial implications associated with real estate transactions, accurate pricing is pertinent for both buyers and sellers (Gunn, *et al***,** 2022).

**Corresponding Author Email: adebisiibrahim97@gmail.com*
**ORCID: https://orcid.org/0000-0002-8797-1369*
**Tel: +2347034845265*

A real estate house price prediction system depends highly on advanced technologies and analytical methodologies to estimate the value of a property based on a particular feature (Pedregosa, *et al***,** 2022). This system incorporates various data sources, including property details, historical sales data, market trends, and socio-economic indicators (Eghagha, 2019). By employing predictive modelling techniques, the system provides an estimate of a property's value, aiding stakeholders in making informed decisions about buying or selling (Rahman *et al*, 2021).

Machine learning has to do with developing algorithms that predicts or make decisions without recourse to writing pseudocodes by making the algorithms master patterns from a set of data. As far as real estate is concerned, machine learning enables pseudocodes to explore and make analysis of several databases and recognize patterns and relationships between property attributes and their corresponding prices. This enables the system to generate accurate predictions that adapt to changing market conditions (Brank *et al*, 2011).

The human instinct to acquire property originated in Roman law and Greek philosophy. Property evaluation is believed to have started in England in the sixteenth century as a result of quest for farming land. Literature on surveying started appearing and while UK preferred "surveying", North America adopted the term "appraisal" (Alvik, 2018). Natural laws, which is also called universal laws, which influenced the concept of private property, became an item of discussion among authors in relation to property theory and international relations when considering protecting private properties and foreign investments 1400s and 1500s (Thamarai *et al*, 2020). The Louisiana Purchase Treaty, signed in 1803, as a result of Louisiana, is considered one of the high-profile deals in the history of real estate. This was when what is now known as Louisiana, was bought for fifteen million from France **(**Zemlyanskiy, 2021). The first real estate brokerage firm, Baird and Warner, formerly known as L. D. Olmsted & Co. was founded in 1855 (Klaasen, 2018)**.** This was followed by the National Association of Realtors in 1908 in Chicago and later renamed National Association of Real Estate Boards in 1916, coinciding with the birth of the term "realtor" for the identification of professionals in real estate.

Housing is a broad term referring to peoples' physical dwelling places which may be in form of apartments or physical structures for lodging or emergency accommodations. Accessibility of affordable, safe and stable housing is paramount to personal health, safety and well-being of people. It equally has effects on the education, employment healthcare as well as social networks of individuals as it influences economic, social and cultural opportunities. This also ensures access to decent housing with the aid of authorities, departments and ministries in charge of housing.

Housing can be broadly categorized into marketing and non-marketing housing. Marketing housing comprises of residential apartments, and commercial outlets which are owned by corporate bodies, individuals with the prices are dictated by the market forces. On the other hand, non-market housing comprises of cooperative, public and social housing. These are usually under the control of government or non-government firms whose aim is to provide alternate and affordable housing to low-income class at subsidized rates with possibility of rent assistance programs (Alvik, 2018).

In another study different advanced machine learning approaches, XGBoost and LightGBM, were proposed for the predicting houses prices in Kuala Lumpur and their efficiencies were measured using root mean squared error (RMSE) and mean absolute error (MAE). The outcome of the study revealed that the XGBoost model performance better as it generated minimal MAE and RMSE in the house price prediction.

Machine learning, nonlinear statistical models, artificial intelligence and Bayesian optimization, which have been applied in house price prediction problems have also been investigated through the application of regression trees, Gaussian process regression, support vector machine and boosting ensemble. It was observed that boosting ensemble regression trees outperformed all, with Gaussian process regression and support vector machine following in that order. Enhancing ensemble regression trees are therefore considered the best approach to improve the prediction of operational house price in Taiwan (Manasa, 2020).

In another study on data pre-processing involving missing values, encoding categorical variables and standardizing numerical features, the data was split into two, namely training and testing sets in order to assess the model. Decision tree models were built using C4.5 and CART algorithms which recursively categorized the data on the basis of features used in creating the tree-like structure. To improve the model's performance, they employed ensemble methods, such as Decision Tree and Random Forest

which combined multiple decision trees to make more accurate predictions. The study compared the performance of the decision tree models and ensemble methods using evaluation metrics such as MAE and RMSE. It was found that the ensemble methods generally outperformed individual decision trees, indicating the benefits of combining multiple models for better predictions. The method also analysed the feature importance derived from the decision tree models to identify the key factors influencing house prices (Rahman, *et al*, 2021). Random Forest Machine Learning Technique as a means of more effective house pricing prediction than house pricing index (HPI) and comparative analyses of various methods have also been discussed (Shuzlina, *et al*, 2021, Lahmiri *et al*, 2023).

MATERIALS AND METHODS

Data Collection and Exploration: Boston housing dataset gotten from Kaggle was used to develop the model. The Boston Housing dataset is a widely used in real estate research. It contains information about various attributes of houses in Boston area, these include proximity to highways, rate of crime, median property value and the average number of rooms per house. Kaggle serves as a source of data for understanding the factors influencing house prices, it contains various features that influence house prices in the Boston area.

Exploratory data analysis is required to build a regression model, as it helps researchers to gain insight into the data and also discover patterns which help in the selection of relevant features to build the model. Below are the features that are available in the dataset used in this experiment:

CRIM: the per capita rate of crimes by town (which reveals the safety of the environment)

ZN: the proportion of land meant for large residential areas

INDUS: the proportion of land meant for non-retail businesses (serving as an indicator for noise and environmental pollution)

CHAS Binary variable: which reveals proximity to Charles River (serving as an indicator for recreational activities and scenic views)

NOX: reflecting nitric oxides concentration (serving as an indicator for pollution)

RM: the average number of rooms per dwelling (serving as an indicator for house prices for larger living spaces)

AGE: the proportion of older owner-occupied units (serving as an indicator for maintenance needs)

DIS: weighed distances to employment canters (serving as an indicator for proximity to job opportunities and urban amenities)

RAD: index of accessibility to radial highways

TAX: full-value property tax rate (serving as an indicator for property values)

PTRATIO: pupil-teacher ratio (serving as an indicator for quality education)

B: the racial composition (proportion of Black residents, serving as an indicator for historical impacts)

LSTAT: percentage of population in lower socio-economic status

MEDV: median value of owner-occupied homes (target variable), representing the median price of houses in thousands of dollars.

Table 1: Reading Dataset

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PT Ratio	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.58	6.998	45.8	6,0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.50	7.147	54.2	6,0622	3	222	18.7	396.90	5.33	36.2

Table 2: Data Description

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PT Ratio	B	LSTAT	MEDV
Count	486.000	486.000	486.000	486.0000	506.00	506.000	486.000	506.0000	506.0000	506.0000	506.0000	506.0000	486.0000	506.0000
Mean	3.6119	1.2119	11.08200	0.07000	0.5547	6.2846	68.5185	3.7950	9.5494	408.2732	18.4555	356.6740	12.7154	22.5328
Std	8.7202	23.3899	6.8359	0.2553	0.1159	0.7026	27.9995	2.1057	8.7073	168.5371	2.1649	91.2947	7.1558	9.1971
Min	0.0063	0.0000	0.4600	0.0000	0.4850	3.5610	2.9000	1.1296	1.0000	187.0000	12.6000	0.3200	1.7300	5.0000
25%	0.2537	0.0000	5.1900	0.0000	0.4490	5.8855	45.1750	2.1002	4.0000	279.0000	17.4000	375.3775	7.1250	17.0250
50%	3.5603	0.0000	9.6900	0.0000	0.5380	6.2085	76.8000	3.2075	5.0000	330.0000	19.0500	391.4400	11.4300	21.2000
75%		12.5000	18.1000	0.0000	0.6240	6.6235	93.9750	5.1884	24.0000	666.000	20.2000	396.2250	16.9550	25.0000
Max	88.9762	100.0000	27.7400	10.0000	0.8710	8.7800	10.0000	12.1265	24.0000	711.0000	22.0000	396.9000	37.9700	50.0000

The dataset was divided into two, namely the testing and the training sets because the model employs a supervised learning technique; 70% of the dataset was utilized for training, while 30% was utilized to test the model performance.

Data Pre-processing: There were various procedures involved in the data pre-processing for the Boston housing dataset. First, to run summary statistics to get information about all numerical columns in the dataset, secondly, was to check for missing values which was found and was handled using mode imputation. Robust strategies were used to identify outliers and deal with them in order to prevent biased (skewed) predictions. However, the absence of categorical variables made the pre-processing simpler. Additionally, since the dataset already included useful attributes, feature engineering was not required.

Model Development: The model was developed using Extra Tree algorithm which was implemented with the use of Extra-Tree Regressor accessible in Python Scikit-learn (sklearn) machine learning libraries.

It is a supervised classification and regression machine learning technique as well as being an ensemble learning algorithm which constructs multiple decision trees using random subsets of features and random thresholds. This randomness reduces overfitting and enhances model robustness, making it effective for various predictive tasks in machine learning. This model will then be compared to another ensemble machine learning algorithm Random Forest algorithm which is a supervised classification and regression machine learning technique. It uses the concept of ensemble learning to resolve complex problems by incorporating numerous classifiers to improve its accuracy.

Extra Trees and Random Forest are similar as they are both ensemble methods based on decision trees. However, they differ in their level of randomness and the way they select features and thresholds. Extra Trees tend to be faster to train and more robust to noisy data, while Random Forest may yield more interpretable individual trees.

RESULTS AND DISCUSSION

The study conducted a thorough analysis by comparing the predicted values and the actual values of house prices and quantifying the disparities. The comparative data is as shown in Table 1.

Table 3. Actual and Predicted Values and their Differences

S/N	Actual Value	Predicted Value	Difference
1	28.873	29.652	-0.779
2	27.097	27.951	-0.854
3	20.024	20.236	-0.212
4	20.935	21.109	-0.174
5	24.017	20.089	3.928
6	19.927	19.515	0.412
7	28.325	28.409	-0.084
8	18.389	19.187	-0.798
9	21.264	19.865	1.399
10	23.028	23.863	-0.835
11	26.771	26.777	-0.006
12	31.706	32.044	-0.338
13	20.478	20.492	-0.014
14	20.610	20.140	0.47

The findings unveiled that, while precise predictions weren't achieved in every instance, the variations between the predicted and actual values generally fell within a relatively narrow band of ±5. This reveals a notable level of precision in the predictions, affirming the efficacy of the model. Even in cases where exact values weren't matched, the model demonstrated a commendable degree of proximity to the actual prices, which further validates its reliability for real estate house price estimations.

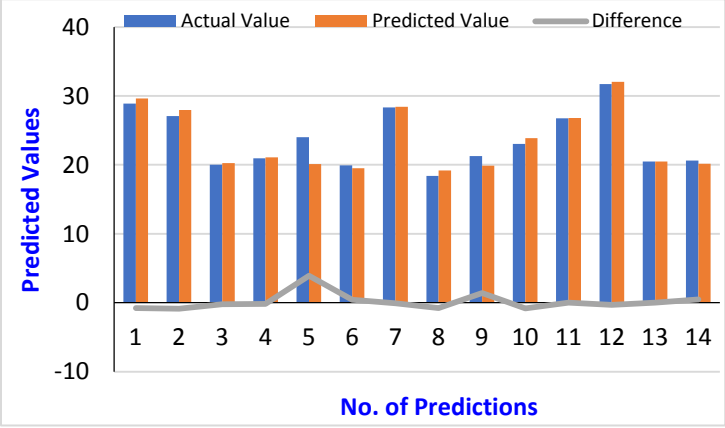

Fig 1: Comparison of Actual and Predicted Values and their differences

Performance Evaluation: The performance evaluation metrics for the prediction models on the test set were performed using Mean Squared Error (MSE) and Mean Absolute Error (MAE). The MSE and MAE for the random forest model was calculated showing the following results:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$= \frac{1}{152} \times 1290.5434299999993$$
$$= 8.490417$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| = \frac{1}{152} \times 338.23 = 2.225197$$

And the MSE and MAE for the Extra Tree model was calculated and showed:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{152} \times 1137.72975$$
$$= 7.485059$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| = \frac{1}{152} \times 312.129000$$
$$= 2.053480$$

Comparing the Mean Squared Error (MSE) and Mean Absolute Error (MAE), lower values indicate better performance. Hence, an MSE of 7.485059 and MAE of 2.053480 for the proposed model is better than an MSE of 8.490417 and MAE of 225197 for the random forest model. This means that the proposed model using Extra Tree algorithm performed better than the model with Random Forest algorithm.

*Conclusion***:** In this article, we delved into the versatility of machine learning algorithms in generating optimal predictive models. The research resulted in the successful development of a robust model tailored for predicting Real Estate house prices. Additionally, a meticulous comparison between two formidable ensemble machine learning techniques revealed that the Extra Tree algorithm outperformed the Random Forest algorithm in the realm of real estate price prediction. The model exhibited a noteworthy level of accuracy in approximating actual prices, affirming its trustworthiness and efficacy in estimating real estate house prices.

Declaration of Conflict of Interest: The authors declare that there is no conflict of interest among them.

Data Availability: Data are available upon request from first author.

REFERENCES

Chen, J, (June, 2023) What Is Real Estate? Investopedia **www.researchgate.net/publication/40220202.**

Thamarai, M; Malarvizhi. SP (2020) House Price Prediction Modelling Using Machine Learning, *Int. J. Information Engrg. and Elect. Bus.*, 12(2): 15- 20.

Akhil, T; Gayathri, CD; Singh, A; Gnanacharitha, G; Zabeeulla ANM (2023) Real Estate Price Prediction Using Machine Learning. *Int. J. Cur. Sci.* 13(2): 224-237

Zemlyanskiy, O. (2021) *Analysis of the Real Estate Market for Business Activities***.** *Vest. Univ.* 1(1); 60-68

Kuvalekar, A; Manchewar, S; Mahadik, S; Jawale, S. (2020) House Price Forecasting Using Machine Learning. *Proc. of the 3rd Int. Conf. on Adv. in Sci. & Tech.*

Gunn, LD; Saghapour, T; Giles-Corti, B; Turrell, G. (2022). Exploring inequities in housing affordability through an analysis of walkability and house prices by neighbourhood socioeconomic disadvantage**.** *Cities & Health*, 6(3): 1-19.

Pedregosa, F; Varoquaux, G; Gramfort, A; Michel, V; Thirion, B; Grisel, O; Duchesnay, E; (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learning*, 12(1): 2825-2830

Eghagha, NW (2019) A Spatial Disparity Analysis of Fire Station Distribution in Lagos, Nigeria. *Int. J. Res. Sci. Inno.* 6(7): 58-66

Rahman, S; Zulkifley, N; Ibrahim, I; and Mutalib, S. (2021) Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur**.** *Int. J. of Adv. Comp. Sci. and App.* 12(12): 736-745.

Brank, J; Mladenić, D; Grobelnik, M; Liu, H; Flach, PA; Garriga, GC; and Toivonen, H. (2011). Feature Selection. In C. Sammut & G. I. Webb

(Eds.), *Ency. of Mach. Learning* 1(1): 402–406. US. doi:10.1007/978-0-387-30164-8_306

Alvik, I. (2018) Protection of Private Property in the Early Law of Nations**.** *J. of the Hist. of Int. Law*, 20(2): 220

Library of Congress Research Guides. *Louisiana Purch. Pri. Doc. in American Hist.* Archived from the original on June 25, 2019. Retrieved May 18, (2022)

Klaasen, RL (2018) Brief History of Real Estate Appraisal and Organizations. *Appr. J.*, 44(3): 376–381

Manasa, J; Gupta, P. (2020) Machine Learning based Predicting House Prices using Regression Techniques. *2nd Int. Conf. on Inno. Mech. for Ind. App.*: 2-63

Shuzlina, AR; Zulkifley, NH; Ibrahim, I; Mutalib, S (2021) Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur *Int. J. of Adv. Comp.r Sci. and Appls*, 12(12): 1-13

Lahmiri, S; Beckiros, S; Avdoulas, C. (2023) A Comparative Assessment of Machine Learning Methods for Predicting House Prices Using Bayesia Optimization**.** *Dec. Anal. J.* 6(1): 1-14