Machine learning techniques for house price prediction: A literature review

Preprint · December 2023

DOI: 10.13140/RG.2.2.31544.52480

1 author:

Ritu Ritu
Bournemouth University
1 PUBLICATION **0** CITATIONS

SEE PROFILE

Machine learning techniques for house price prediction: A literature review

Ritu
dept. computing and informatics
Bournemouth University
Bournemouth, UK
https://orcid.org/0009-0000-9874-5753

***Abstract*— This paper explores the machine learning techniques such as Random Forest, Stacked Generalization Regression, Linear Regression and XGBoost for house price prediction proposed by different researchers in their literatures. The study discusses twenty literatures which are focused on solving house price prediction problem with the help of machine learning and published in last five years only. The models proposed in these literatures are compared on basis of their accuracy, performance metrices used for evaluation and the dataset on which the model is implemented. This review finds the importance of data variables available in a dataset and revels that the performance metrices like root-mean-square error (RMSE), R-squared score, relative root mean squared logarithmic error (RSMLE), mean absolute error (MAE) are important parameters for evaluation of house price prediction model. The results reveal the superior performance of models in different hypotheses, providing insights into their effectiveness for accurate house price prediction. This paper provides the latest and up-to-date information gathered after reviewing the existing research papers for house price prediction based on machine learning and will act as a guide those who are interested in this field.**

Keywords— machine learning, house price prediction, review

I. INTRODUCTION

The real estate market is essential to a city's economic framework, with housing accessibility, affordability, and prosperity as critical success indicators. Residential properties are one of the most important property market segments in developed countries [1]. Knowing the exact value of a property is very important for money matters and making rules about how cities grow. Many groups, like property owners, buyers, banks, and government bodies, are affected by property values, which have a big impact on how cities work [2]. With increase in population the demand for houses is also increasing which results rise in cost of houses. Another reason for the growth in price of houses is the dependency on the real estate agents which consider the necessity of people as business and try to gain high profits. The agent-based model in real estate market results loss to both buyer and seller which results in very unorganised structure of this industry. To optimize the public and private investments in real estate market there is requirement to organize this industry, which needs a model that will help to decide the price of a house or building. Real estate price prediction models rely on two types of variables: non-geographical factors that detail the building's features, and geographical elements such as its distance from landmarks, nearby services, and specific geographical coordinates [3]. All these individual features consist of class of features which makes it a complex problem. The complexity of the problem rises while considering the trend of the house price, as it converts the problem to a time series problem [4]. Several researchers propose artificial intelligence (AI) based models to tackle this challenge and offer an effective solution for accurately predicting prices. This literature review paper aims to discuss the research findings related to the prediction of house prices using machine learning (ML). The objective is to point, dissect, and assess the major ideas, patterns, and gaps in the field. Along with that, this paper provides the understanding of the current state-of-art machine learning models and insights of the existing literatures, to highlighting potential areas for future research.

- The following are the primary contributions of this paper:
- Investigating AI-based machine learning algorithms discussed in existing literatures.
- Evaluating the approaches followed in these literatures.
- Discussing the potential, challenges, and directions for using ML algorithms to address the to address the complexities of real estate industry.
- Providing a review of the use of AI-based ML techniques in real estate prediction and forecasting.

The approach used in this paper is explained in the following section. Following that, the third section includes a review of the literature. The fourth portion is a thorough discussion that includes detailed assessments of the included literature. Finally, the final part summarises the main findings and makes recommendations for prospective future study projects.

II. METHODOLOGY

To achieve the objective of this review paper journals databases are utilized, using specific keywords to search the relevant literatures. In literature analysis, the model proposed in the research is discussed and categorized based on the specific machine learning techniques utilized, enabling a structured and systematic evaluation. Then approaches are compared on basis of datasets used, evaluation metrics applied, and the key findings derived from these studies. Then the key findings and emerging trends observed throughout the reviewed literature are discussed. Finally, the conclusion will highlight the main outcomes of the paper. Furthermore, potential avenues for future research will be proposed, aiming to guide and inspire further exploration in this field. The flowchart of methodology is presented in Fig. 1.

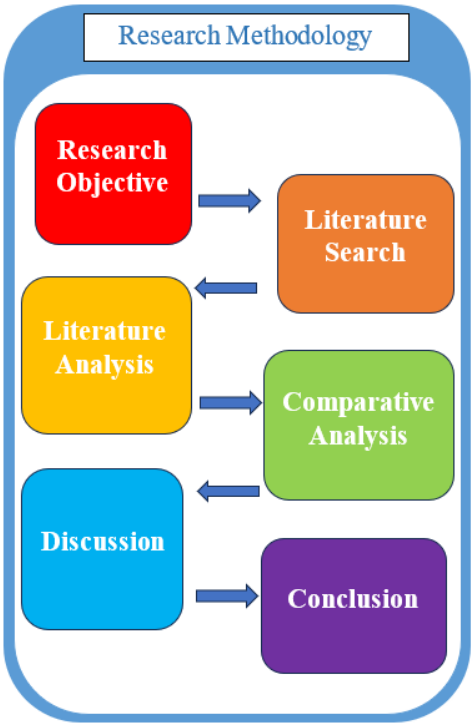

Fig. 1. Flowchart -- Research Methodology

III. Literature Review

The field of house price prediction has seen an increase in research efforts, with various machine learning models being used to improve accuracy and efficiency. In this literature review, we explore and compare findings from several studies, each proposing different models and methodologies for predicting house prices.

The Random Forest model consistently outperforms other house price prediction algorithms in many studies. For example, study [5] shows that it outperforms Decision Tree and Linear Regression approaches in terms of root-mean-square error (RMSE) (3.316) and variance (0.628) in price prediction. Another study [6], which used apartment transaction data from Gangnam, Korea (from 2006 to 2017), supports the claim of earlier discussed study, naming the Random Forest model as the most accurate when compared to a traditional hedonic pricing model. A study [7] on Chengdu dataset compares the Random Forest model to XGBoost, Support Vector Regression, a back-propagation network, and a regionally weighted regression model. The Random Forest model outperforms competitors with a mean absolute error of 1517.29 and best r2 of 0.83 after 5-fold cross-validation. In another study [8], Random forest model has even outperformed deep learning model. The model is evaluated against neural network model, inverse distance weighting and Kriging and for this evaluation accuracy metrices such that mean absolute error (MAE), RMSE, mean absolute error percentage (MAPE) and mean absolute scaled error (MASE) are used and the model is implemented on the dataset taken form Ministry of Land, Infrastructure, and Transport (MOLIT) which has sales record from 2016 to 2019, Seoul, South Korea. Meanwhile, a study [9] conducted on Greater Sydney dataset demonstrates the superiority of two models, Random Forest and GWR, over Linear Regression, Support Vector Regression, tree-based methods, Gradient Boosting-based approaches, and sequential neural networks. Contrastingly, several studies have also investigated the performance of linear and non-linear machine learning models in the domain of house price prediction models for example, A study [10] by Sagala and Cendriawan proposes a machine learning model that uses Linear Regression to accurately predict house prices. This model produced lowest RMSE of 0.0334 and a smallest R value of 0.7 using the Maribisnis dataset from 2014-2015. Expanding on the effectiveness of linear models, another study [11] confirms Linear Regression's dominance in terms of accuracy. When compared to Lasso Regression and Decision Tree, this model achieved an impressive accuracy of 83.54% while using real estate data from the Pune city portal. However, a non-linear machine learning approach presented in paper [12] by Calainho, Minne and Francke outperforms the linear model. Based on dataset of real estate transactions in New York from 2000 to 2019, this study shows that non-linear models outperform linear models, particularly Ordinary Least Squares (OLS). The non-linear model not only outperforms in terms of accuracy, but it is also subjected to rigorous testing, including 5-fold cross-validation and hyperparameter adjustment, to improve its prediction powers.

In recent study [13] on house price prediction models, an XGBoost model is introduced by Grybauskas which outperforms 15 machine learning methods, including CatBoost Classifier, Random Forest Classifier, and SVM. The XGBoost model achieves an outstanding accuracy of 0.002% using data from Vilninus city (18,992 properties) from May to August 2020. Similarly, in a study [14] by Karamanou Kalampokis and Tarabanis, a machine learning model applying the XGBoost algorithm appears to be the best performer in terms of accuracy, using data from the Scottish data portal. Another study [15] compares Extreme Gradient Boosting (XG-Boosting) with a hedonic regression model. The XGBoost model outperforms with an accuracy of 84.1%, whereas the hedonic regression model only achieves 42%. In a study [16] by Lahmiri a boosting ensemble regression model was found to be the most accurate when compared to support vector regression and Gaussian process regression. Using the dataset obtained from the Taiwan Ministry of the Interior between June 2012 and May 2013, the model is validated with 10-fold cross validation to confirming its stability and low prediction error. Despite these accomplishments, it is important to recognize the limitations. Some research, such as [17], present better Gradient Boosting (Stacking algorithm) and compare it to various models. However, the problem of not considering the account key macroeconomic and microeconomic elements for property price evaluation is observed. In a larger sense, a study [18] by Garcia, Lopez and Sanchez investigates ensemble learning methods based on boosting and bagging and compares them to a linear regression model. The dataset, which was obtained from a real estate portal in a Spanish city, adds to the landscape of house price prediction models. These all research collectively contribute to a better knowledge of the advantages, disadvantages, and nuances of using XGBoost and boosting ensemble regression models to predict housing prices.

Various studies in the field of housing price prediction models have suggested different models, each claiming superiority in terms of accuracy. On the California Housing Dataset, a study [19] proposes a Hybrid LGBM model that outperforms LGBM, XGBoost, AdaBoost, and GBM. The hybrid LGBM and XGBoost models had the lowest MSE,

MAE, and MAPE values of 0.193, 0.285, and 0.156, suggesting a significant accuracy benefit. The study [20] offers a Hybrid Model that incorporates linear regression, clustering analysis, nearest neighbour classification, and Support Vector Regression (SVR). This model outperforms its competitors on datasets from Istanbul, Turkey, and KAGGLE Ames Housing, with superior accuracy measures such as RMSE, MAE, MAPE, and R2. A Hybrid Bayesian Optimization Stacking model is given and tested against 18 machine learning models in a study [21] by Zhan in terms of RMSE, the HBOS-CatBoost model outperforms the HBOB-XGBoost and HBOT-ConvLSTM models. Another study [22] offers another viewpoint, claiming that the Stacked Generalization Regression model surpasses other models in accuracy despite its high temporal complexity. This study compares standard and sophisticated machine learning approaches to Random Forest, XGBoost, and LightGBM, reaching an excellent relative root mean squared logarithmic error (RSMLE) of 0.16350 on the testing set. And study [23] by Alvarez, Rangel and Montiel investigates incremental learning, offering a model that beats SVR in terms of absolute error and R2. The model, which was tested on a dataset from Austin's open data portal, demonstrates the power of incremental learning in improving accuracy. In conclusion, this diverse research provides useful insights into the growing landscape of house price prediction models, each offering distinct advantages and considering different techniques to improve accuracy.

IV. DISCUSSION

This literature review reveals different machine learning techniques proposed by other researchers to deal with the challenge of house price prediction. Each discussed model has its own strengths and weaknesses. Random forest model emerges as a top performer in many literatures where the model is evaluated on different kind of datasets. Stacked Generalization model also has shown good accuracy, but this model has a significant time complexity because of that it may not be ideal to implement the model on real world dataset where the data is complex and in large volume. Linear Regression model has also shown good performance in some studies whereas non-linear models have provided better performance. Boosting and Bagging algorithms like XGBoost has also provided good results in some papers, which are also improved by using ensemble techniques. It is observed that the combination of boosting and bagging algorithms contributes to enhanced predictive capabilities, with XGBoost particularly standing out in terms of regularization, handling spare data, and achieving impressive accuracy rates. Numerous research works present hybrid models that integrate various machine learning techniques or incorporate sophisticated approaches like Bayesian Optimization. These hybrid methods frequently produce better outcomes, suggesting that model development may still be innovative. The variety of datasets—from real estate data from Pune, Gangnam, and New York to housing datasets from Beijing and Boston—highlights the necessity for models to adjust to various socioeconomic and geographic situations. Model creation and evaluation require a sophisticated approach due to the diverse characteristics of the dataset.

After reviewing all the literatures, it is observed that developing a machine learning model for house price prediction requires a proper methodology. The common strategy observed in discussed literatures starts with discussing the problem and collecting the relevant dataset. Then, the data analysis is performed on the dataset to develop data understanding which helps in deciding the parameters for data preprocessing, where data cleaning is performed, and valuable features are extracted. The clean data is split into training and testing dataset to train and evaluate the model. The dataset is implemented on the machine learning models. In many papers effectiveness of the model is evaluated using cross validation method. Also, the comparison of model with other state-of-the-art models is a common practice to find an optimal model.

Selection of right evaluation metrices is important to check the performance of a machine learning model. As observed from the TABLE. 1, which is providing the information about evaluation methods used in different literatures, the most common used metrices are MAPE, RMSE, R-square. These are the metrices which are almost used in all the studies. MAPE is mean absolute error percentage, as name signifies this metrices deals with the error percentage and RMSE is root mean square deviation which measures the average difference between predicted and the actual value. In a few papers, it is observed that many metrices are used for comparison purposes.

In summary, the literature review gives a systematic process for constructing and assessing machine learning models that are effective for predicting house prices, in addition to offering insights into these models. Predictive analytics in real estate may progress as a result of future research that examines hybrid models, takes into account additional evaluation criteria, and tackles implementation issues in the real world.

V. CONCLUSION

To sum up, research indicates that machine learning models are a reliable method for predicting housing prices. When it comes to accuracy measures, the Random Forest model shows itself to be a reliable option, routinely surpassing other models. Stacking Generalization Regression is accurate, but its computational cost should be considered. In some situations, linear regression is still competitive, especially when paired with hybrid strategies. The Ensemble Learning and XGBoost models show good predictive power. The optimal model selection is contingent upon the features of the dataset and the prioritised evaluation measures. To improve the precision and applicability of house price prediction models, more investigation into hybrid models, optimization strategies, and evaluation of diverse datasets are suggested. This literature review showcases the variety of methods and their corresponding performances, offering a view of machine learning models for housing price prediction. The conversation contributes to the developing field of real estate predictive analytics by highlighting similarities, difficulties, and potential research areas.

TABLE I. EVALUATION MATRIX OF MODELS USED IN LITERATURES

Title	Evaluated Against	Evaluation Metrices
House Price Prediction System using Machine Learning Algorithms and Visualization	Decision Tree and Linear Regression	B. root-mean-square and r-squared score
Housing Price Prediction via Improved Machine Learning Techniques	Random Forest, XGBoost, and LightGBM, Hybrid Regression and Stacked Generalization Regression	k- fold validation technique is used to find an acceptable bias-variance trade off.
Real Estate Price Prediction using Supervised Learning	Linear Regression, Lasso Regression, and Decision Tree	RMSE, MAE, RSE, MSE, and R-square
A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea	conventional hedonic pricing model	MAPE, coefficient of dispersion (COD), and R-squared
Real estate market price prediction model of Istanbul	linear regression, polynomial regression, decision trees, random forests, and XGBoost	MAPE, R-square
Analysis and Evaluation of Housing Price for Chengdu Urban	XGBoost, Random Forest, Support vector regression, back-propagation network, and geographically weighted regression	5-fold cross-validation technique is used to evaluate the models
A Comparative Study of Machine Learning and Spatial Interpolation Methods for Predicting House Prices	This proposed model evaluated against neural networks, inverse distance weighting, and kriging	MAE, RMSE, MAPE, and MASE
A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate	This proposed model compared with linear models (OLS)	Chained Paasche index (CP), RMSE, RW and the Expanding Window (EW).
Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia	Linear regression, support vector regression, tree-based methods, Gradient Boosting-based approaches, and sequential neural network, OLS, Lasso, and Ridge	R-square, MAPE and MAE
Incremental learning for property price estimation using location-based services and open data	The proposed model evaluated against SVR.	Absolute error, R-square
Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic	15 machine learning models	Accuracy metrices
A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization	Support vector regression, and Gaussian process regression	MAE, RMSE, MARE, MAPE
A hybrid machine learning framework for forecasting house price	HBOS-XGBoost, HBOS-CatBoost, HBOS-AdaBoost, HBOB-XGBoost, HBOB-LightGBM, HBOB-BLS, HBOT-based, HBOT-CNN, HBOT-ConvLSTM, BPNN, CNN, GRU, LSTM, BLS, Seq2Seq-GRU, Seq2Seq-LSTM, Deep Transformer, AdaBoost, CatBoost, GBDT, ExtraTrees, LightGBM, XGBoost, Bagging, KNN, RF, SVR	MSE, RMSE, MAE, MAD, MAPE, RMSLE, EVS, ME, MPD, MGD and PL.
A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices	The proposed model compared with LGBM, XGBoost, AdaBoost and GBM	MSE, MAE, MAPE
A Novel Hybrid House Price Prediction Model	The proposed model compared with multiple linear regression, Lasso, ridge regression, Support Vector Regression (SVR), AdaBoost, decision tree, random forest and XGBoost regression	RMSE, MAE, MAPE, and R-square
Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times	Linear regression model.	R-square
House price prediction using hedonic pricing model and machine learning techniques	The model is compared with hedonic regression model	MAE, RMSE, Recall, Precision, F-measure, Sensitivity
House Price Prediction with An Improved Stack Approach	Linear, Lasso, AdaBoost, Elastic Net, Ridge, Random Forest.	RMSE

REFERENCES

[1] S. Jabin, M. S. Suhi, Md. F. Arefin, and K. Md. Hasib, "Comparison of Different Sentiment Analysis Techniques for Bangla Reviews," in 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC), IEEE, Sep. 2022, pp. 288–293. doi: 10.1109/R10-HTC54060.2022.9929495.

[2] Q. Gao, V. Shi, C. Pettit, and H. Han, "Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia," Land use policy, vol. 123, p. 106409, Dec. 2022, doi: 10.1016/j.landusepol.2022.106409.

[3] F. Alvarez, E. Roman-Rangel, and L. V. Montiel, "Incremental learning for property price estimation using location-based services and open data," Eng Appl Artif Intell, vol. 107, p. 104513, Jan. 2022, doi: 10.1016/j.engappai.2021.104513.

[4] W. Guang and S. Zubao, "Research on the Application of Integrated RG-LSTM Model in House Price Prediction," in 2023 IEEE 5th International Conference on Power, Intelligent Computing and Systems (ICPICS), IEEE, Jul. 2023, pp. 348–353. doi: 10.1109/ICPICS58376.2023.10235649.

[5] M. S. Supriya, G. S. Vinayak, V. R. Patgar, and V. Mahajan, "House Price Prediction System using Machine Learning Algorithms and Visualization," in 2023 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), IEEE, Jul. 2023, pp. 1–6. doi: 10.1109/CONECCT57959.2023.10234749.

[6] J. Hong, H. Choi, and W. Kim, "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, vol. 24, no. 3, pp. 140–152, Feb. 2020, doi: 10.3846/ijspm.2020.11544.

[7] M.-J. Li et al., "Analysis and Evaluation of Housing Price for Chengdu Urban," in 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE, Dec. 2020, pp. 175–178. doi: 10.1109/ICCWAMTIP51612.2020.9317395.

[8] J. Kim, Y. Lee, M.-H. Lee, and S.-Y. Hong, "A Comparative Study of Machine Learning and Spatial Interpolation Methods for Predicting House Prices," Sustainability, vol. 14, no. 15, p. 9056, Jul. 2022, doi: 10.3390/su14159056.

[9] Q. Gao, V. Shi, C. Pettit, and H. Han, "Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia," Land use policy, vol. 123, p. 106409, Dec. 2022, doi: 10.1016/j.landusepol.2022.106409.

[10] N. T. M. Sagala and L. H. Cendriawan, "House Price Prediction Using Linier Regression," in 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), IEEE, Jul. 2022, pp. 1–5. doi: 10.1109/ICCED56140.2022.10010684.

[11] V. Matey, N. Chauhan, A. Mahale, V. Bhistannavar, and A. Shitole, "Real Estate Price Prediction using Supervised Learning," in 2022 IEEE Pune Section International Conference (PuneCon), IEEE, Dec. 2022, pp. 1–5. doi: 10.1109/PuneCon55413.2022.10014818.

[12] F. D. Calainho, A. M. van de Minne, and M. K. Francke, "A Machine Learning Approach to Price Indices: Applications in Commercial Real Estate," The Journal of Real Estate Finance and Economics, Apr. 2022, doi: 10.1007/s11146-022-09893-1.

[13] A. Grybauskas, V. Pilinkienė, and A. Stundžienė, "Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic," J Big Data, vol. 8, no. 1, p. 105, Dec. 2021, doi: 10.1186/s40537-021-00476-0.

[14] A. Karamanou, E. Kalampokis, and K. Tarabanis, "Linked Open Government Data to Predict and Explain House Prices: The Case of Scottish Statistics Portal," Big Data Research, vol. 30, p. 100355, Nov. 2022, doi: 10.1016/j.bdr.2022.100355.

[15] J. Zaki, A. Nayyar, S. Dalal, and Z. H. Ali, "House price prediction using hedonic pricing model and machine learning techniques," Concurr Comput, vol. 34, no. 27, Dec. 2022, doi: 10.1002/cpe.7342.

[16] S. Lahmiri, S. Bekiros, and C. Avdoulas, "A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization," Decision Analytics Journal, vol. 6, p. 100166, Mar. 2023, doi: 10.1016/j.dajour.2023.100166.

[17] H. Zhang, K. Wang, M. Li, X. He, and R. Zhang, "House Price Prediction with An Improved Stack Approach," J Phys Conf Ser, vol. 1693, no. 1, p. 012062, Dec. 2020, doi: 10.1088/1742-6596/1693/1/012062.

[18] R.-T. Mora-Garcia, M.-F. Cespedes-Lopez, and V. R. Perez-Sanchez, "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times," Land (Basel), vol. 11, no. 11, p. 2100, Nov. 2022, doi: 10.3390/land11112100.

[19] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," Engineering Reports, vol. 5, no. 4, Apr. 2023, doi: 10.1002/eng2.12599.

[20] S. Özöğür Akyüz, B. Eygi Erdogan, Ö. Yıldız, and P. Karadayı Ataş, "A Novel Hybrid House Price Prediction Model," Comput Econ, vol. 62, no. 3, pp. 1215–1232, Oct. 2023, doi: 10.1007/s10614-022-10298-8.

[21] C. Zhan, Y. Liu, Z. Wu, M. Zhao, and T. W. S. Chow, "A hybrid machine learning framework for forecasting house price," Expert Syst Appl, vol. 233, p. 120981, Dec. 2023, doi: 10.1016/j.eswa.2023.120981.

[22] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing Price Prediction via Improved Machine Learning Techniques," Procedia Comput Sci, vol. 174, pp. 433–442, 2020, doi: 10.1016/j.procs.2020.06.111.

[23] F. Alvarez, E. Roman-Rangel, and L. V. Montiel, "Incremental learning for property price estimation using location-based services and open data," Eng Appl Artif Intell, vol. 107, p. 104513, Jan. 2022, doi: 10.1016/j.engappai.2021.104513.