Mundo Fesc E-ISSN 2216-0353 P-ISSN 2216-0388P-ISSN 2216-0388 **Artículo Original**

Mejora de la predicción de precios de vivienda mediante modelos de aprendizaje automático de conjunto: un estudio comparativo con datos de vivienda de EE. UU.

Enhancing House Price Prediction Using Ensemble Machine Learning Models: A Comparative Study with U.S. Housing Data

Recibido: 03 de septiembre de 2025
Aprobado: 21 de diciembre de 2025

Forma de citar: A. D. Gbadebo and A. O. . Ayodele, "Mejora de la predicción de precios de vivienda mediante modelos de aprendizaje automático combinados: un estudio comparativo con datos del mercado inmobiliario estadounidense", Mundo Fesc, vol. 16, no. 34, pp.501-522 Jan. 2026, doi: 10.61799/2216-0388.1885.

Abiola Olaide Ayodele
ayodeleabiola1994@gmail.com
0009-0000-9366-8613,
Department of Information Technology,
University of Ilorin
Nigeria

Adedeji Daniel Gbadebo
agbadebo@wsu.ac.za
0000-0002-1929-3291
Department of Accounting Science
Walter Sisulu University
Mthatha, Eastern
Cape, South Africa.

*Autor para correspondencia:
agbadebo@wsu.ac.za

© 2026. Fundación de Estudios Superiores Comfanorte.

Mejora de la predicción de precios de vivienda mediante modelos de aprendizaje automático de conjunto: un estudio comparativo con datos de vivienda de EE. UU.

Resumen

Antecedentes: Uno de los principales desafíos en el mercado inmobiliario, especialmente cuando las condiciones políticas y económicas no son estables, es la predicción precisa de los precios de las viviendas. Los modelos de pronóstico deben ser flexibles para afrontar la creciente complejidad del mercado, pero esta flexibilidad es difícil de lograr con los métodos de predicción convencionales. Objetivo: El objetivo de este trabajo es comparar el poder predictivo de cinco modelos de aprendizaje automático: XGBoost, LightGBM, CatBoost, HistGradientBoosting y Bayesian Ridge Regression, mediante un análisis del mercado inmobiliario estadounidense en Kaggle, para identificar el modelo con mejor rendimiento y las características más relevantes. Métodos: Se utilizó una muestra de 1460 muestras y 32 variables del período 1991-2024. La efectividad de los modelos se evaluó mediante medidas de regresión y clasificación. Se realizaron análisis de importancia de las características para determinar las variables que siempre influyeron en las predicciones. Resultados: El mejor rendimiento se obtuvo con CatBoost, con una precisión de entrenamiento de 0,94, un R^2 de prueba de 0,86 y un MCC de 0,81. OverallQual y GrLivArea fueron las variables que demostraron una alta influencia. Los modelos de conjunto CatBoost y XGBoost superaron a Bayesian Ridge en no linealidades y variables categóricas. Medidas de clasificación como la especificidad y el coeficiente Kappa de Cohen también aportaron información sobre la robustez del modelo. Conclusiones: Los modelos propuestos ofrecen resultados competitivos o superiores en comparación con la literatura previa y, además, mejoran la interpretabilidad. Los resultados resaltan la importancia de las técnicas de aprendizaje de conjunto y el análisis multidimensional en la creación de modelos de predicción inmobiliaria robustos y basados en evidencia. Los datos de vivienda de EE. UU. utilizados en este proyecto se descargaron de Kaggle e incluyen todas las características necesarias para predecir correctamente el precio de las viviendas, ya que se trata de datos históricos que abarcan el período de 1991 a 2024.

Palabras clave: Preprocesamiento de datos, Ensemble, Selección de características, Pronóstico de precios de vivienda, Aprendizaje automático.

Enhancing House Price Prediction Using Ensemble Machine Learning Models: A Comparative Study with U.S. Housing Data

Abstract

Background: one of the major challenges in the house market, particularly when political and economic conditions are not stable is proper prediction of house prices. Forecasting models need to be flexible to face rising market complexities but the flexibility is hard to realise with the conventional prediction methods. Purpose: The purpose of the paper is to compare the predictive power of five models of machine learning, which are XGBoost, LightGBM, CatBoost, HistGradientBoosting, and Bayesian Ridge Regression with the help of a Kaggle U.S. housing analysis to outline the best-working model and the most meaningful features. Methods: A sample of 1,460 samples and 32 variables of 1991 to 2024 were utilized. The effectiveness of the models was assessed using both the regression and classification measures. Importance analyses of features were carried out in order to determine variables that always affected the predictions. Results: The best performance was achieved with CatBoost, training accuracy is 0.94, test R 2 = 0.86 and MCC = 0.81. OverallQual and GrLivArea were all the variables that were found to be highly influential. CatBoost and XGBoost ensemble models performed better than Bayesian Ridge in non-linearities and categorical variables. Classification measures such as specificity and Cohen Kappa also gave more information about the robustness of the model. Conclusions: The models proposed result in the competitive or superior outcomes in comparison to the previous literature and, in addition, enhance the interpretability level. The results highlight the significance of the ensemble learning techniques and the multidimensional analysis in the creation of strong and evidence-based real estate prediction models. The U.S. housing data utilized in the current project was downloaded at Kaggle and includes all of the features required to correctly predict the price of houses since it is historical data covering the years 1991 to 2024.

Keywords: Data pre-processing, Ensemble, Feature selection, House price forecasting, Machine learning.

Introduction

The house price prediction is a critical key factor in the property sector. The property business in the last few years has seen many bizarre patterns, most of which have been caused by politics, economics as well as social factors. Models that predict the market are important in the decisions of investors, home owners, and policy makers because it is based on the predictions that their decisions are made on investment, urban planning and the funding strategy are made. Preferably, housing market forecasts are supposed to be reliable just as weather forecasts. Nonetheless, the history of designing effective and reliable forecasting models has been difficult and many methods have been used that have been proved to be unprofitable as time goes by.

The professionals working in the real estate business are majorly dependent on the past and the present information to make accurate guesses, and even the most expert analysts cannot be sure of what lies ahead. Although it is a given truth that there is a certain degree of uncertainty, expert analyses can be of great help. The S&P CoreLogic Case-Shiller Home Price Index reported that the United States recorded a home price growth of 4.1% per year in January 2025, the slightest higher than the 3.9% per year growth in December 2024, based on the index of single-family home values. Zillow data showed that a typical home in February 2025 is valued at some $357,000. A buyer who put a 20 percent down payment on such property and financed the rest at the market average mortgage rate of 6.76 percent on a 30-year fixed mortgage would be making a monthly principal and interest payment of $1,854. These figures point to the problem of affordability and quality price estimation in financial planning.

It is challenging to get nearly 100 percent accuracy in house price prediction despite many researches that have been performed over the years. To illustrate, [1] put predictive modeling on Vietnamese housing market and estimated the contribution of attributes like floor area, quality of neighborhood, and age of buildings to the cost of properties. As shown in [2], Ordinary Least Squares (OLS) have been proved to be useful in both the developed and the developing world in understanding the housing market. On the same note, [3] used OLS to establish the determinants of the housing prices in the United States that are of significance with a more focus on the structural and neighborhood aspects of the neighborhood. The importance of several variables like the number of rooms, the age of the property, and closeness to city centres was further indicated through multiple linear regression using the Boston housing data [4], with the values of Spearman correlation coefficients. These conventional statistical tools provide useful information, but fail to work in the measurement of complex and non-linear relationships which are common in the housing markets.

More modern research has examined the alternative modeling models. As an example, [5] compared ARIMA and Vector AutoRegression (VAR) models in order to forecast prices of houses in Texas. Both models were effective in capturing upward trends though VAR has given extra information about the effect of the household income and employment factors on the price changes. CMA process, which is reported in [6], is another common method of determining the value of property through the issue of recent sales of close properties. Its efficiency, however is very much reliant on the accessibility of correct and updated market data.

This study uses five advanced machine learning models: XGBoost, LightGBM, CatBoost, HistGradientBoosting and Bayesian Regression to surpass the drawbacks of conventional models as well as to enhance prediction accuracy. The models were chosen because they have better predictive potential and are capable of addressing the non-linear relationship between housing characteristics. The machine learning methods are especially effective in the real estate market, where the datasets are huge and heterogeneous and their dependence is complex. These models can make stronger predictions than traditional statistical methodologies by picking up trends using historical data.

Lastly, this study is well structured and divided into five sections. Section Two is the methodology section, which describes how the data was collected, pre-processed, and engineered with features as well as the machine learning models. Section three provides the result and discussion section which includes the model assessment metrics, comparison of performance and discussion of results. Section Four shows how the study can be applied in practice, what implications it has on policy and what are the limitations of the study. Section Five is the conclusion of the paper that summarizes the most important contributions, essential findings, and recommendations to be made in further research, including the possibilities of incorporating real-time data streams and newly developed predictive technologies.

Focusing on the significance of feature engineering, ensemble learning, and the close attention to the choice of predictive models, the given study can contribute to the overall understanding of housing price forecasting and offer a credible framework to the stakeholders in the discussed real estate market; Section Four will be section that made the conclusion clear, Section Five will be section that talked of acknowledgement and Section Six will be section that talked about references.

Materials

Some researchers have proved the benefits of machine learning in the house prices forecasting. As an illustration, [7] examined the real estate data of Seoul, South Korea, involving 20,000 records and attained a level of R-squared of 0.83. [8] studied more than 1 million home listings on Zillow in the United States achieving an R-squared of 0.87. A dataset-based study of 10 million property records [9] showed that even simply based decision-tree methods obtained high accuracy of prediction and had an R-squared of 0.86. The use of geographical data also enhanced the performance of the models as indicated in [10] where more than 15 million properties were dealt with to represent the spatial impacts on the housing prices.

The engineering of features that are innovative is especially crucial in the setting where the information is insufficient or incomplete. Some studies may be dedicated to the incorporation of other sources of real-time information, including social media sentiment, transportation development, and macroeconomic changes to enhance the accuracy of predictions even further and offer the in-depth insights into the market dynamics up to the same accuracy. Some of the existing prediction strategies are also reliable as the outcome of a R-squared of 0.87 when [11] analyzed 5 million records is reported. On the other hand, [12]. employed a more traditional statistical procedure to forecast the prices of houses in Lagos, Nigeria and achieved a rather satisfactory R-squared of 0.67 [13] focused on the data in the United States and proved that the advanced methods of analysis provide the R-squared value of 0.83. In [14] and [15]. Conducted research about the house prices in Portugal and achieved important results of 0.74 and 0.85 respectively.

In [16], she researched the application of various classification algorithms in the prediction of the resale prices of houses. The tools used in analyzing the data in the study were linear regression, decision trees, K-means clustering and random forests to ascertain the impacts of such factors as physical features, location and the economic statuses on the prices of houses. In order to establish the optimum behavior of these algorithms, the study employed the Root Mean Square Error (RMSE) as its primary metric that will be deployed to decide the best model to employ in the prediction of the value of resale. [17] proposed a regression hybrid model to enhance the accuracy of house prices. Their design fitted Lasso and Gradient Boosting regression models, and it is important to point out that innovative feature engineering is strongly suggested, especially when the data is limited. The significant role in deriving significant patterns using small datasets is essential in feature engineering where new input variables are generated or the existing ones are transformed to enhance model performance.

Predictive performance can be significantly enhanced by transforming the existing variables or introducing new features, e.g., the distance to the transport, the school rating, the neighborhood socio-economic index, the level of property taxes, etc. In [18], according to a comparative study in the UK, a modern feature-based engineering approach performed better than old-fashioned approaches with an R-squared of 0.75. In different studies, the range of R-squared values of house price prediction is between 0.67 and 0.91, which means that the level of prediction is different based on data used, approach, and market conditions. Although high R-squared values indicate the quality of the predictors, other measurements in evaluation such as MAE and RMSE are also very important to determine the overall performance of the model.

Ensemble learning has become an effective method of enhancing prediction strength. Ensemble methods decrease variance and bias by combining a number of base learners thus enhancing the generalization performance [26]. The overall performance is directly related to the stability and diversity of base classifiers [27]. It is always observed that ensemble methods perform better than individual predictors in estimating real estate prices, as they have greater precision and utility [28]. As an illustration, [29] used integrated model adaptive sampling (IMAS) model, but oversampling was used to predict multiclass housing prices with an accuracy of 78%. Gradient boosting methods and random forest algorithms have also turned out to be useful. In America, the House Price Index was predicted using random forests with 5% margin of error [30]. On the same note, decision trees, gradient boosting and random forests performed better than multiple linear regression models in Australia [31].

Predictive capabilities have also been increased by artificial neural networks (ANNs). Research reports indicate that it has been successfully used in China (China Real Estate Index System) [33], Lagos, Nigeria [34], [35], Boston, United States [36] and Taranto, Italy [37]. ANNs are able to model the complex non-linear relationships between housing characteristics and market dynamics thus being applicable to a range of housing markets and large-scale datasets. The set of machine learning models, feature engineering, and ensemble approaches is a big advancement as compared to the traditional approaches to statistics. Moreover, [10] included any geographic data in their analysis of a survey of more than 15 million of houses and emphasized the role of spatial variables on the house prices and enhanced the model accuracy.

Other than these, machine learning models have demonstrated certain strengths in numerous situations. An example case is XGBoost which has been applied to predict the price of houses in London utilizing the characteristics of the property, the crime rates in the neighborhood, and the school rating with a resulting R-squared of 0.89. LightGBM has been shown to be computationally efficient and also accurate on high dimensional data like housing information in Tokyo whereby it obtained an R-squared of 0.88 and also took shorter times to train as compared to other boosting models. It has been applied with Categorical features in CatBoost which is quite good at modeling housing prices in India, including features on the nature of the property, the neighborhood and nearby infrastructure, and has been shown to be highly competitive against other prediction models. In its histogram-based method, HistGradientBoosting has been useful with large scale datasets such as the U.S. Housing Price Index data, and it can use a smaller amount of memory with a corresponding R-squared of 0.87. Bayesian regression has the interpretability advantage that enables probabilistic forecasting, and quantification of uncertainty, which is of crucial importance in making decisions as a property investor.

These models can be connected with each other in one predictive system to obtain a more holistic picture of the housing market. Instead, they are able to find non-linear correlations between property characteristics, macroeconomic variables, and local facilities. Additionally, with the ability to predict using models together with ensemble methods, further increases in the accuracy can be realized, which would offer strong and effective advice to the buyers, as well as the policymakers. The increasing access to real-time information, like transaction logs, demographic fluctuations, and geographic signals, will allow retraining the model continuously, which will keep the predictions current and precise enough to act on.

Conclusively, forecasting the prices of houses with absolute accuracy can be elusive since the market is volatile, but using highly developed machine learning models can greatly enhance a higher level of prediction accuracy. The models do not only reflect intricate trend in property data, but also introduce other outside forces like the location, economic variables and past trends. With the housing markets constantly changing, the combination of well analytics with judgment is bound to be the pillar of sound, practical housing market forecasting.

According to [18], a comparative analysis of two estimation techniques in the UK housing market established that the more contemporary one greatly outdid the older, which gave an R-squared value of 0.75. Overall, the findings of these experiments are that the predictability of house prices is not very similar as the values of R-squared lie between 0.67 and 0.91. Although large values of R-squared may indicate better model performance, it is also crucial to pay attention to the other metrics of evaluation, including mean absolute error (MAE) and root mean squared error (RMSE), which would allow to understand the predictive performance of the model in general.

Real estate is not merely a key driver of economic development, but it is among the most important social issues. The housing affordability and the property value trend is growing in the interest of investors, policymakers as well as the general population especially in the market that is undergoing a rapid urbanization and population growth. The real estate market has become a very competitive and at times in transparent market, where profitability of prices is essential to make the right decisions [19]. However, with time, some other methods of predicting the price of property are established with the aim of considering both the intrinsic factors like the size, age, and amenities of the property, and the extrinsic factors like the location of the property, infrastructure and neighborhood features [20]. Availability of good and timely data is vital in the selection and training predictive models that can work based on changing market condition.

Our approach analyzes these parameters that the customer or investor sets to arrive at the most efficient pricing approaches according to their requirements and goals [21]. Prices of properties are subject to change depending on a variety of different factors such as location, area, amenities available and structural differences. It has taken scientists in the last ten years to develop and test predictive models that will accurately estimate these prices [22]. It has been found that the literature reviews contain a large variety of approaches such as single-model predictions, and hybrid models which combine many algorithms in order to enhance the accuracy [23]. Numerous factors determine housing prices and some of these factors include time, geographical location, the age of the property, conditions of the neighborhoods, community services and accessibility of transportation. The traditional single-predictor models are not very effective in the noisy data, which constrains their predictive capabilities [24].

In a bid to come up with the best algorithms that are useful in predicting the housing market, scholars have examined past rates of growth or the house price indexes, which in most cases is determined based on the average prices of houses in a country [25]. Ensemble learning has become a very efficient method in this regard. Ensemble methods help to improve generalization as they aggregate a number of base learners, which reduce the variance and bias of a single model [26]. Ensemble learning is able to minimize prediction errors and enhance stability by combining the output of different classifiers. The quality and diversity of the learner make up an ensemble model very important in defining the overall performance of the ensemble model [27].

Some works have emphasized the effectiveness of team models over individual models. As an illustration, an ensemble learning algorithm used to predict the prices of real estate properties performed better than a single decision tree in both flexibility and accuracy and has a high degree of practical utility in the analysis of markets [28]. On the same note, [29] also used an integrated model adaptive sampling (IMAS) method and oversampling in prediction of multiclass house prices, which gained an accuracy rate of 78%. This approach enhanced the precision of forecasting through addressing the imbalance of classes that are usually difficult to address when dealing with heterogeneous housing data. Other models based on decision tree are random forests and gradient boosting models, which have been found to be good at predicting. Prediction of in [30] was done using a random forest model. United States House Price Index at 5 percent error level. Decision trees, gradient boosting and random forest algorithms when applied to the Australian housing market have continued to perform better than the multiple linear regression models [31].

ANNs have as well been effective in forecasting the prices of houses in various geographic areas. It was used successfully in studies in China (China Real Estate Index System) [33], Lagos, Nigeria [34], [35], Boston, United States [36], and Taranto, Italy [37]. The models are able to represent complicated and non-linear correlations among housing attributes and market fashions and are, therefore, very applicable with huge and dynamic datasets. These methods have a high level of international applicability, which emphasizes the importance of using sophisticated machine learning algorithms with situational market intelligence to better predict accuracy. Testing of the dataset using four various regression algorithms which are Lasso Regression, Logistic Regression, Decision Tree, and Support Vector Regression is one of the approaches put forward in the paper [38]. In the comparison of the error measurements of such algorithms as R-Squared Value, Mean Absolute Error, Mean Squared Error and Root Mean Squared error, Decision Tree has proved to be the best algorithm with a high level of accuracy of86.4% and low levels of error as opposed to Lasso Regression that has a low level of accuracy of 60.32.

Most of our efforts regarding ensemble learning in our predictive model have been largely released to be a success in the enhancement of predictions [39]. Where ensemble models may differ as required in the data needs and prediction way where even a modicum of betterment can cause a significant difference. This combination of an ensemble of two or more ensemble members is determined by the form of integration the developer would consider appropriate to the data i.e. Constant weighting Functions and Non- Constant weighting Functions. Another methodology has been adopted in the article [40]. The paper has concentrated on the association between different studies connected to the analysis of housing market price. Much attention is paid to hedonic price modeling and its application to the house price market and the potential presence of a submarket.

[41] employs the Decision Tree to make predictions following the greatest accuracy in terms of prediction values among various other algorithms run namely Linear Regression, Multiple Linear Regression, Decision Tree Regressor and KNN. Besides other parameters such as the no. of bedrooms, carpet area, built up area, age of the property, zip code, no. of bathroom in the property, latitude and longitude of the property, they have added two more features to the prediction as air quality and the crime rate. The other proposed system employed Lasso and Random Forest regression models and selected the most appropriate model to the data based on the values of error [42]. The information was forwarded to 6 steps such as data pre-processing, 50:50 split of data, training of the data under the Lasso and Random Forest models and testing on the test data, and selecting the best model. [43] uses the UiPath Studio Platform in order to build the RPA Flowchart. The UiPath Studio offers the data scraping using the help of scraping wizards. A set of machine learning algorithms are compared and applied on the data. Comparison of the boosting algorithms is conducted namely XGBoost, Light BGM and CatBoost. Random Forest was discovered to perform well when dealing with small quantities of data and it is not enhanced by increasing the number of samples. CatBoost was declared the obvious victor in contests. RPA was a significant advancement in regard to efficient extraction and would not tend to create errors [44].

It compares various classification algorithms like Decision Tree, Logistic Regression, and random Forest, AdaBoost, and Naive Bayes with an accuracy of 92, 81.5, 86.5, 96, and 88 respectively. AdaBoost and Decision Tree with C 5.0 were chosen that predicts house values and with help of rules, profit or loss were predicted. Separate analysis of various machine learning algorithms which include Multiple Linear Regression, Elastic Net Regression, Ridge Regression, Ada Boosting Regression, LASSO Regression and Gradient Boosting has been studied using a publicly available output data of a given area within the USA [45]. The scores obtained of the algorithms were 0.73, 0.66, 0.73, 0.78, 0.73 and 0.97 respectively. Gradient Boosting was the most appropriate algorithm because it produced small error values.

Methodology

USA house pricing dataset was considered for this project, it was downloaded from kaggle site which contains important features for making accurate house price prediction. The historical housing data was obtained from Kaggle from 1991 through 2024. Before preprocessing should take place, analyzing the dataset is very crucial in order to have a better understanding of the data and its features. This analysis is premised on the accessing of correlation matrix to determined relationship between various features. In correlation analysis, +1 and -1 correlation mean positive (Linear relationship) and negative (Non-Linear relationship) which are independent variables. Thus, analyzing the variables helps to understand the interrelation between the features (Independents variable) and the target (dependent variable). This is done for the purpose of identifying how one variable influences the others. Moreover, another importance is to make an informed decision. Figure 1 show correlation between variables.

Data preprocessing before training in machine learning involves cleaning the collected data. This can be done by removing any form of outliers that can cause models' bias, encoding object data into numeric; this is done when dealing with data in classification into regression and normalizing the numeric features. In this study, some specific features as mentioned above were considered as independent variables, this is determined by some specific criteria which are very important to predict house price. Moreso, the diagram below illustrates the data processing" life cycle.

The study employed several supervised learning algorithms to model historical United States "housing price data, with the primary objective of predicting property sale prices using key structural and quality-related predictors. The dependent variable was the sale price, while the independent variables included LotArea, YearBuilt, TotalBasementSF, GarageCars, GrLiveArea, and OverallQuality. To ensure robust evaluation, the dataset was randomly divided into an 80% training subset and a 20% testing subset, preserving the statistical distribution of the original dataset. Model training was conducted using XGBoost, LightGBM, CatBoost, HistGradientBoosting, and Bayesian Ridge Regression. GridSearch-based hyperparameter optimization was applied to each model to improve learning efficiency and generalization performance, taking into account model-specific considerations such as learning rate, tree depth, regularization strength, sampling ratios, and leaf-wise growth strategies. Because these boosting-based models are computationally intensive, hyperparameter ranges were adjusted to balance model complexity and training time.

Each of the predictive models was constructed according to its formal mathematical structure. XGBoost optimized a regularized objective function expressed as:

$$L(\phi) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

with regularization defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| w\|^2$$

and with the second-order Taylor approximation of the objective:

$$L^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t)$$

LightGBM, using histogram-based gradient boosting, selected splits based on loss reduction computed as:

$$\Delta L_{split} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma$$

CatBoost employed ordered boosting, minimizing:

$$L^{(t)} \approx \sum_{i=1}^{n}\left[g_i f_t^{(ordered)}(x_i) + \frac{1}{2}h_i\left(f_t^{(ordered)}(x_i)\right)^2\right] + \Omega(f_t)$$

with categorical target statistics computed as:

$$CTR(x) = \frac{\sum y + prior}{count + prior\ weight}$$

HistGradientBoosting minimized a histogram-based loss expressed as:

$$L^{(t)} \approx \sum_{j=1}^{b} \left(G_j w_j + \frac{1}{2} H_j w_j^2 \right) + \Omega(w)$$

Bayesian Ridge Regression assumed a model of the form:

$$y = Xw + \varepsilon, \quad \varepsilon \sim N\left(0, \alpha^{-1} I\right)$$

with a prior: $p(w|\lambda) = N(0, \lambda^{-1} I)$, posterior: $p(w|X, y, \alpha, \lambda) = N(\mu_w, \Sigma_w)$ and $\Sigma_w = (\lambda I + \alpha X^T X)^{-1}$; $\mu_w = \alpha \Sigma_w X^T y$

with hyperparameters determined by maximizing the marginal likelihood:

$$logp(y|X, \alpha, \lambda) = -\frac{n}{2} log(2\pi) - \frac{1}{2} log\left|\alpha^{-1} XX^T + \lambda^{-1} I\right| - \frac{1}{2} y^T \left(\alpha^{-1} XX^T + \lambda^{-1} I\right)^{-1} y \quad (9)$$

The methods of modelling, optimization processes, and measures of evaluation offered a strict and in-depth structure to examine the predictive behaviour of the contemporary machine learning logic in making a forecast of the housing prices based on organized real estate information.

The evaluation of the model performance was done on held-out test set to find out the accuracy of the fitted models to capture the structural patterns of the housing market. Multiple regression metrics were used to check the accuracy of the forecast and the distribution of error. Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Coefficient of Determination (R 2) and Explained Variance Score (EVS) are calculated. The performance measures in the form of confusion matrices were calculated with the use of TP, TN, FP, and FN and Accuracy, Precision, Recall, Specificity, Matthews Correlation Coefficient (MCC) and Cohen Kappa (^{2}) were estimated.

Result and Discussion

The results of the present research are in high consistency with the general literature on the usefulness of advanced ensemble and boosting algorithms in predicting house prices in different markets. Previous studies have always indicated that ensemble models outdo single estimators because they are able to capture non-linearity, lower variance, and increase the rate of generalization. To explain, in the literature of predicting real estate, the study of [28] demonstrated that the use of ensemble learning is superior to any single decision-tree model in terms of both adaptability and accuracy. Equally, the ensemble model of the IMAS that is presented in [29] realized an impressive accuracy rate of 78 percent when oversampling was applied to address the issue of class imbalance, which can be taken to show that data preprocessing and intelligent sampling methods can contribute to a great extent of predictive modeling capability in heterogeneous housing data.

This is also visible in the current study: CatBoost, XGBoost and LightGBM performed better than Bayesian Ridge Regression, which serves to prove the superiority of the ensemble-based gradient boosting models in processing the mixed character of features, non-linear dynamics, and high-dimensional interactions. Specifically, CatBoost showed a high level of predictive stability and a high training accuracy of 0.94 and R 2 of 0.86, which is close to the ranges of the previous studies. The use of gradient-boosting-based models has often been known to outperform the rest of the models, as evidenced by [45] who achieved the highest score (0.97) with Gradient Boosting as opposed to other models such as Elastic Net (0.66) and Lasso Regression (0.73). These previous data are further supported in the current study because CatBoost showed a strong performance on the problem presented with not only high accuracy, but also feature importance patterns that are highly interpretable.

Remarkably, the performance metrics of this study also indicate the competitive or even better results in comparison with various previous studies. As an example, [7] was able to replicate a 20,000 observation in Seoul with an R 2 of 0.83, whereas [8] with more than a million Zillow listings was able to achieve an R 2 of 0.87. The CatBoost model of the current study achieved the R 2 of 0.86 - placing it at the highest levels of results obtained in the literature, even though it was built with a smaller dataset and with a more complicated classification-regression hybrid assessment. In addition, XGBoost and LightGBM models provided higher R 2 of 0.86 and 0.85, respectively, generated better confusion matrices with low false-positive and false-negative rates, and showed better discriminative ability than previous models that had concentrated on the regression error metrics alone.

Another unique aspect of this work is that it combines sophisticated indicators of classification namely MCC, specificity, and Cohen Kappa alongside traditional regression indicators (MAE, RMSE, MAPE). Although the error margins and R2 have been considered as primary indicators of inaccuracy in earlier studies, this study uses a multi dimensional analysis. This holistic has added rigor to performance comparison. As an illustration, CatBoost had a perfect specificity of (1.00), hence, the model was very good at making correct predictions on low priced houses without making a false prediction. This level of specificity is immensely uncommon in the housing prices research literature, since previous models, e.g., [40], [41] and [42], were primarily concerned with the variance and error reduction, but not with classification strength.

In addition, there is a great degree of congruency with the results found in the previous empirical literature in the pattern of feature importance. In line with [9], Kuvalekar et al. and other large researches like the ones by Case and Quigley, the variables of OverallQual and GrLivArea emerged as the most powerful predictors of sale price. These characteristics have consistently been confirmed as important housing value determinants since it takes into consideration both the structural quality and the living space. The implication of CatBoost and XGBoost modeling these features with high fidelity is that both of these algorithms are capable of implicitly encoding domain knowledge. This is compared to more basic and more linear models, like Bayesian Ridge, which failed to entirely accommodate non-linear interactions leading to poorer generalization ($R^2 = 0.84$) and a little larger errors (MAE = 0.12).

Other attributes like GarageCars, Year Built, and TotalBsmtSF were also significant contributors to model performance though marginalized returns were observed, which is better addressed by ensemble learners as compared to linear models. This is complementary to the results in [30] and [31], which showed the tree-based models to be more effective in the representation of non-linear dynamics and interaction among the structural housing characteristics. Equally, the identification of time and geographical variables as proposed in [30], [31] and [40] further increased model interpretability as the boosting algorithms were able to distinguish between older and newly built houses, various neighborhoods, and finer market divisions.

In comparison to other research studies like [33] and [37] where Artificial Neural Networks were implemented in international markets (China, Boston, Lagos, Taranto), the current findings indicate that boosting algorithms can be as successful or even more successful than ANNs in addition to the ability to provide better explainability and less training complexity. This is of great importance particularly in real estate forecasting where model transparency is of great importance to the stakeholders in the industry [46].

Lastly, the overall evaluation findings, including the training accuracy of up to 0.94, the test accuracy of around 0.90, the R^2 of between 0.84 and 0.86, and the best MCC and specificity results, show that the models developed in this study do not only fit within the upper range of performance reported in the previous literature but also make a methodological contribution to the problem, which has traditionally been payed significant attention to through regression. The confusion matrices and interpretive analysis of misclassifications in prediction give a better perspective of the model behavior and patterns of price segmentation, which is not easy to deepen in literature.

Altogether, the findings discussed in this paper, when compared to existing studies, demonstrate that modern boosting algorithms and especially CatBoost can provide the latest performance in the domain of housing price prediction. High level of performance measures, good feature engineering and the use of a sound model architecture are some of the features that make this work a step forward in the context of previous predictive studies that only used normal regression methods or constrained evaluation frameworks. The findings validate the hypothesis that using ensemble learners in conjunction with meticulously selected housing characteristics results in very precise, trustworthy, and interpretable predictions of house prices, which provides a solid argument of the practical utility of these designs in real estate analytics.

Conclusions

The findings of the present paper indicate the groundbreaking characteristics of created machine learning models in the task of housing price prediction with unprecedented rates of accuracy and predictability. CatBoost, XGBoost and LightGBM turned out to be the best performing algorithms since they produced a better performance in training and testing data when it comes to regression and classification performances. These results indicate the power of the existing algorithms to approximate the nonlinear relationships, which are present in the markets of real estate.

Under these predictive tools, the policy-makers, financial institutions, and other real estate stakeholders can make more informed and data-based decisions that will result in greater efficiency in the markets, affordability of housing, and lessening financial risks. The fact that machine learning is introduced to the process of housing valuation is quite a significant move towards modernizing the realm of the real estate business, thus, making it more transparent, more just, and responsive to the market forces.

In the meantime the machine learning must be implemented into the housing markets with models of a good regulation measures that will ensure fairness, responsibility, and privacy of the personal information. Ethical concerns, model governance, and engagement by the people have to be parts of any of the changes of policy concerning the usage of algorithms to make decisions in the real estate business.

Overall, this paper demonstrates that machine learning has the potential to deliver important sources of innovation to the housing market analysis and policy-making. However, they can be realized only due to the assistance of the strategic investment in the data infrastructure and constant monitoring of the model, regulatory control, and cooperation across the sector. The analysis can also be performed in future to widen the scope of socioeconomic, environmental, and infrastructural issues incorporated in predictive models to get a clearer picture of the housing markets.

Based on the way that the machine learning models in question might be accurate in the future, it is implied that high-level algorithms, including CatBoost, XGBoost, and LightGBM, can be used to predict the housing market by the policymakers, real estate professionals, and financial institutions. The models were very accurate and dependable in the fact that they were able to contribute immensely to the valuation systems of the property. Governments should adopt and use automated valuation models (AVMs) to advance fairness, transparency, and efficiency in the real estate and tax assessment.

It is postulated that proper policies ought to be taken to hold machine learning models utilized in predicting houses accountable. These policies should require a mechanism of explainability, fairness and accountability, and minimisation of biases and privacy of the homeowners. To assure the social confidence and protect the financial systems against the errors of the models, the regulatory authorities are also expected to set minimum performance strategies of the predictive models used in the official activity of the housing market.

Finally, the stakeholders should establish periodically monitoring and validation systems to determine the present performance of the predictive models. Since housing markets are dynamic, it is extremely critical to update the models regularly and re-calibrate them with the latest information to ensure that the rates of prediction accuracy might not fall. The models would be able to adapt to the economic cycles, market shocks, and demographics through continuous improvement models.

References

[1] T. T. Nguyen, T. L. Tran, and Q. T. Bui, "The hedonic pricing model applied to the housing market in Vietnam," ResearchGate, 2020. [Online]. Available: https://www.researchgate.net/publication/343551323

[2] S. Malpezzi, "Economic analysis of housing markets in developing and transition economies," in Handbook of Regional and Urban Economics, vol. 3, P. C. Cheshire and E. S. Mills, Eds. Elsevier, 1999, ch. 44, pp. 1791–1864. [Online]. Available: https://ideas.repec.org/h/eee/regchp/3-44.html

[3] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via improved machine learning techniques," Procedia Computer Science, vol. 174, pp. 433–442, Jan. 2020. [Online]. Available: https://doi.org/10.1016/j.procs.2020.06.111

[4] Q. Zhang, "Housing price prediction based on multiple linear regression," Scientific Programming, vol. 2021, no. 3, pp. 1–9, Oct. 2021. [Online]. Available: https://doi.org/10.1155/2021/7678931

[5] Z. Wu, "Time series forecasting of Texas housing prices: A comparison between the ARIMA and VAR models," Theoretical and Natural Science, vol. 80, pp. 20–27, 2025. [Online]. Available: https://doi.org/10.54254/2753-8818/2025.GL19918

[6] Investopedia, "How much is my home worth?," 2021. [Online]. Available: https://www.investopedia.com/how- much-is-my-home-worth-5213913

[7] M. Thamarai and S. P. Malarvizhi, "House price prediction modeling using machine learning," International Journal of Information Engineering and Electronic Business, vol. 12, no. 2, pp. 15–20, Apr. 2020. [Online]. Available: https://doi.org/10.5815/ijieeb.2020.02.03

[8] S. Juneja, N. Chaudhary, R. Gupta, O. Kaushik, M. Ishan, and A. Sharma, "House price prediction using machine learning algorithms," International Journal for Research in Applied Science and Engineering Technology, 2023. [Online]. Available: https://doi.org/0.22214/ijraset.2023.54259

[9] A. Kuvalekar, S. Manchewar, S. Mahadik, and S. Jawale, "House price forecasting using machine learning," in Proc. 3rd Int. Conf. Advances in Science & Technology (ICAST), Apr. 2020. [Online]. Available: https://ssrn.com/abstract=3565512. https://doi.org/10.2139/ssrn.3565512

[10] A. Kaushal and A. Shankar, "House price prediction using multiple linear regression," in Proc. Int. Conf. Innovative Computing & Communication (ICICC), 2021. [Online]. Available: https://ssrn.com/abstract=3833734 or http://dx.doi.org/10.2139/ssrn.3833734

[11] N. H. Zulkifley, S. A. Rahman, U. N. Hasbiah, and I. Ibrahim, "House price prediction using a machine learning model: A survey of literature," International Journal of Modern Education and Computer Science, vol. 12, no. 6, pp. 46–54, Dec. 2020. [Online]. Available: https://doi.org/10.5815/ijmecs.2020.06.04

[12] O. Adetunji et al., "Prediction of house prices in Lagos-Nigeria using machine learning models," 2021.

[13] M. Al-Saidi et al., "House price prediction using machine learning algorithms," 2020.

[14] Z. Li, "A comparative study of regression models for housing price prediction," pp. 810–816, Aug. 2024. [Online]. Available: https://doi.org/10.62051/qjs7y352

[15] R. Annamoradnejad and I. Annamoradnejad, "Machine learning for housing price prediction," Oct. 2022. [Online]. Available: https://doi.org/10.4018/978-1-7998-9220-5.ch163

[16] P. Durganjali and M. V. Pujitha, "House resale price prediction using classification algorithms," in Proc. Int. Conf. Smart Systems and Services (ICSSS), Mar. 2019. [Online]. Available: https://doi.org/10.1109/ICSSS.2019.8882842

[17] S. Juneja, "House price prediction using machine learning algorithms," International Journal for Research in Applied Science and Engineering Technology, vol. 11, no. 6, pp. 3156–3164, Jun. 2023. [Online]. Available: https://doi.org/10.22214/ijraset.2023.54259

[18] S. Lu, Z. Li, Z. Qin, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," in Proc. IEEE Int. Conf. Industrial Engineering and Engineering Management (IEEM), Dec. 2017. [Online]. Available: https://doi.org/10.1109/IEEM.2017.8289904

[19] M. Zaidi et al., "Comparison of linear regression and random forest models for house price prediction in the UK," 2021. Also see Q. Zhang, "Housing price prediction based on multiple linear regression," Scientific Programming, 2021. [Online]. Available: https://doi.org/10.1155/2021/7678931

[20] L. Bork and S. V. Møller, "Forecasting house prices in the 50 states using dynamic model averaging and dynamic model selection," International Journal of Forecasting, vol. 31, no. 1, pp. 63–78, 2015. [Online]. Available: https://doi.org/10.1016/j.ijforecast.2014.05.005

[21] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," Expert Systems with Applications, vol. 42, no. 6, pp. 2928–2934, 2015. [Online]. Available: https://doi.org/10.1016/j.eswa.2014.11.040

[22] V. Plakandaras, R. Gupta, P. Gogas, and T. Papadimitriou, "Forecasting the US real house price index," Economic Modelling, vol. 45, pp. 259–267, 2015. [Online]. Available: https://doi.org/10.1016/j.econmod.2014.10.050

[23] S. V. Boyapati, M. S. Karthik, K. Subrahmanyam, and B. R. Reddy, "An analysis of house price prediction using ensemble learning algorithms," Research Reports on Computer Science, May 2023. [Online]. Available: https://doi.org/10.37256/rrcs.2320232639

[24] J. J. Jui et al., "Flat price prediction using linear and random forest regression based on machine learning techniques," in Lecture Notes in Electrical Engineering, vol. 678, Springer, 2020, pp. 205–217. [Online]. Available: https://doi.org/10.1007/978-981-15-6025-5²19

[25] S. Sun, "Real estate price prediction based on data mining," Modern Electronic Technique, vol. 40, no. 5, pp. 126–129, 2017.

[26] S. V. Boyapati, M. S. Karthik, K. Subrahmanyam, and B. R. Reddy, "An analysis of house price prediction using ensemble learning algorithms," Research Reports on Computer Science, May 2023. [Online]. Available: https://doi.org/10.37256/rrcs.2320232639

[27] B. Cao and B. Yang, "Research on ensemble learning-based housing price prediction model," Big Geospatial Data and Data Science, vol. 1, no. 1, p. 18, 2018. [Online]. Available: https://dx.doi.org/10.23977/bgdds.2018.11001

[28] E. L. Glaeser and C. G. Nathanson, "An extrapolative model of house price dynamics," Journal of Financial Economics, vol. 126, no. 1, pp. 147–170, 2017. [Online]. Available: https://doi.org/10.1016/j.jfineco.2017.06.012

[29] U. Rajan, A. Seru, and V. Vig, "The failure of models that predict failure: Distance, incentives, and defaults," Journal of Financial Economics, vol. 115, no. 2, pp. 237–260, 2015. [Online]. Available: https://doi.org/10.1016/j.jfineco.2014.09.012

[30] Y. Li, P. Branco, and H. Zhang, "Imbalanced multimodal attention-based system for multiclass house price prediction," Mathematics, vol. 11, p. 113, 2022.

[31] A. B. Adetunji, A. F. Alaba, A. Ajala, N. O. Akande, et al., "House price prediction using random forest machine learning technique," Procedia Computer Science, vol. 199, Feb. 2022. [Online]. Available: https://doi.org/10.1016/j.procs.2022.01.100

[32] A. Soltani, M. Heydari, F. Aghaei, and C. J. Pettit, "Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms," Cities, vol. 131, no. 4, p. 103941, Dec. 2022. [Online]. Available: https://doi.org/10.1016/j.cities.2022.103941

[33] X. Xu and Y. Zhang, "House price forecasting with neural networks," International Journal of Intelligent Systems and Applications, 2021. [Online]. Available: https://doi.org/10.1016/j.iswa.2021.200052

[34] R. B. Abidoye and A. P. C. Chan, "Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network," Pacific Rim Property Research Journal, vol. 24, no. 7, pp. 1–13, Feb. 2018. [Online]. Available: https://doi.org/10.1080/14445921.2018.1436306

[35] A. Al Bataineh and D. Kaur, "A comparative study of different curve fitting algorithms in artificial neural network using housing dataset," in Proc. IEEE National Aerospace and Electronics Conference (NAECON), Jul. 2018. [Online]. Available: https://doi.org/10.1109/NAECON.2018.8556738

[36] M. Ottomanelli, V. Chiarazzo, M. Marinelli, and L. Caggiani, "A neural network-based model for real estate price estimation considering environmental quality of property location," Transportation Research Procedia, vol. 3, pp. 810–817, Dec. 2014. [Online]. Available: https://doi.org/10.1016/j.trpro.2014.10.067

[37] A. Azadeh, M. Sheikhalishahi, and A. Boostani, "A flexible neuro-fuzzy approach for improvement of seasonal housing price estimation in uncertain and non-linear environments," South African Journal of Economics, vol. 82, no. 4, Jun. 2014. [Online]. Available: https://doi.org/10.1111/saje.12047

[38] P. Sobana, M. Balakumaran, S. Bharathkumar, J. Harish, et al., "House price prediction using machine learning," Nov. 2024. [Online]. Available: https://doi.org/10.1201/9781003559085-121

[39] J. M. Moreira, C. Soares, A. M. Jorge, and J. F. de Sousa, "Ensemble approaches for regression: A survey," ACM Computing Surveys, vol. 45, no. 1, pp. 10:1–10:40, Nov. 2012. [Online]. Available: https://doi.org/10.1145/2379776.2379786

[40] Y. Garud, H. Vispute, N. Bisai, and M. Nashipudimath, "Housing price prediction using machine learning," International Research Journal of Engineering and Technology (IRJET), 2020.

[41] P. Patil, D. Shah, H. Rajput, and J. Chheda, "House price prediction using machine learning and RPA," International Research Journal of Engineering and Technology (IRJET), 2020.

[42] A. Kuvalekar, S. Manchewar, S. Mahadik, and S. Jawale, "House price forecasting using machine learning," in Proc. 3rd Int. Conf. Advances in Science & Technology (ICAST), Apr. 2020. [Online]. Available: https://ssrn.com/abstract=3565512; https://doi.org/10.2139/ssrn.3565512

[43] Z. Han, J. Gao, H. Sun, R. Liu, C. Huang, L. Kong, and H. Qi, "An ensemble learning-based model for classification of insincere question," in Proc. Forum for Information Retrieval Evaluation (FIRE), 2019.

[44] A. Varma, A. Sarma, S. N. Doshi, and R. Nair, "House price prediction using machine learning and neural networks," in Proc. IEEE Int. Conf. Innovative Computing and Communication Technology (ICICCT), Apr. 2018. [Online]. Available: https://0.1109/ICICCT.2018.8473231

[45] P. Durganjali and M. V. Pujitha, "House resale price prediction using classification algorithms," in Proc. IEEE Int. Conf. Smart Systems and Services (ICSSS), Mar. 2019. [Online]. Available: https://10.1109/ICSSS.2019.8882842

[46] R. M. Chandra, G. Anuradha, and M. V. Pujitha, "House price prediction using regression techniques: A comparative study," in Proc. IEEE Int. Conf. Smart Systems and Services (ICSSS), Mar. 2019. [Online]. Available: https://10.1109/ICSSS.2019.8882834