A Comparative Study of Urban House Price Prediction using Machine Learning Algorithms

Lale EL Mouna[1], *Hassan Silkan*[1], *Youssef Haynf* [2], *Mohamedade Farouk Nann* [3], *Stéphane C. K. Tekouabou* [4,5]

[1]Laroseri Laboratory, Chouaib Doukkali University, Morocco
[2]The National School of Business and Management of Dakhla, Ibn Zohr University, Morocco
[3]Scientific Computing, Computer Science and Data Science Research Unit (CSIDS), University of Nouakchott, Mauritania
[4] Center of Urban Systems (CUS), Mohammed VI Polytechnic University (UM6P), Hay Moulay Rachid, 43150 Benguérir, Morocco
[5] Department of Computer and Educational Technology, Higher Teacher Training College (HTTC), University of Yaoundé I, Yaoundé, Cameroon

Abstract. Accurate housing price forecasts are essential for several reasons. First, it allows individuals to make informed decisions about buying or selling real estate and to determine appropriate prices. Secondly, it helps real estate agents and investors make better investment decisions and negotiate contracts more effectively. In addition, housing prices are often an indication of the general state of the economy. A price decrease may indicate an economic recession, while an increase in prices may signal economic growth. In this study, we proposed to address this subject by predicting house prices using machine learning by choosing three types of machine learning: Linear Regression (LN), Random Forest (RF) and GradientBoosting (GB). We tested our models on the Melbourne real estate dataset, which includes 34,857 property sales and 21 features.

Keywords: urban real estate, house price, machine learning, house price prediction.

1. Introduction

Housing is a critical component by which the success of a national economy can be measured. When an economy grows, people migrate from cities to rural areas, which leads to an increase in the urban population. As the urban population increases, the demand for housing also increases. The increase in demand drives up housing prices.

The housing price in general is influenced by several variables. The authors [1] define these variables as physical conditions, design, and location. Physical conditions that can be observed through physical perception include the size of the property, the number of rooms, the size of the kitchen and garage, the availability of landscaped space, the land and building size, and the age of the property. The physical characteristics of a house, such as the size of the structure, the year of building, the number of bedrooms and bathrooms, and other elements that determine the internal characteristics of the house can affect the price of the house [2]. While these terms refer to various marketing tactics utilized by real estate developers to interest target investors. For example, the proximity of a property to hospitals, markets, educational centers, airports, major highways, etc. The location has a significant influence over price of a property. The current price of land is determined by the region. Therefore, it is not only in the interest of tenants, but also in the interest of landlords, analysts, and policy makers, as well as urban and regional planning authorities, to understand housing price patterns and their determinants [3]. A computerized forecasting system can help them in making well-informed decisions about the desirability and timing of buying a property [4,5,6,7].

In recent years, machine learning has made significant advances due to increasing computational power, the availability of large datasets and advances in algorithm development [8,9]. These advances have had a significant impact on several industries, including accurate house price forecasting. Using the power of machine learning, real estate professionals and property owners can now make more informed decisions based on reliable estimates of property values.

In the past, estimating property values relied significantly on human expertise and traditional statistical methods. However, these approaches often failed to account for the complexity and non-linear relationships of housing market data. With the advance of machine learning, the forecasting landscape has changed with the ability to extract precious information from large volumes of data.

The aim of this study is to validate and compare three popular machine learning algorithms (linear regression, random forest, and gradient boosting) in the context of house price forecasting. It is intended to provide insights into the effectiveness and applicability of each algorithm to this specific task by examining the strengths and weaknesses of each algorithm. Following this analysis, the reader will have a thorough understanding of the possibilities and limitations of these models and will be able to choose the most appropriate approach for predicting house prices.

We used the Melbourne housing dataset to evaluate the performance of these three models, and the mean absolute error (MAE), the coefficient of determination (R^2) and the root mean square error (RMSE) were used to measure performance.

Our paper on is organized as follows. In Section 2, we report some related work. In Section 3, we present the methodology used in our paper. In the fourth section, we present the results of the different methods. The last section discusses the results obtained and ends with a conclusion.

2. Related work

Forecasting housing prices is a key issue in real estate, finance, and development. Accurate forecasts of the market value of housing allow buyers, sellers, and investors to make informed decisions. In recent years, machine learning techniques have played an important role in the development of housing price forecasting models. This review provides an overview of research in the field of house price forecasting.

Shinde et al [10] used the machine learning algorithms to build a predictive model of house prices. They used techniques such as logistic regression, support vector regression, Lasso regression and decision tree to develop a predictive model. They used data from 3,000 properties. The R-squared values for logistic regression, SVM, Lasso regression and decision tree were 0.98, 0.96, 0.81 and 0.99 respectively.

Dagar et al [11] considered different machine learning methods for predicting house prices based on certain features. The dataset used was based on Bangalore, India, and included 13,000 records and 9 features. They found that Multivariable Linear Regression model is the best solution to this problem, giving the best results in terms of accuracy and error.

Jha et al [12] used several machine learning algorithms to solve real estate market problems, including logistic regression, random forests, voting classifiers and XGBoost.They combined these algorithms with item coding to build an accurate property sales price prediction model that predicted that the negotiated sales price would be higher or lower than the advertised sales price. To evaluate the performance of the model, they evaluated accuracy, precision, findability, F1 rating and error rate. Of the four machine learning algorithms tested, XGBoost had the best performance and the highest model robustness compared to the other algorithms.

.

Hjort et al [13]. addressed the application of enhanced gradient boosting trees (GBTs) for house price prediction and compared the results of different loss functions. GBTs are widely used machine learning algorithms for regression problems, and the choice of loss function has a significant impact on their prediction accuracy. They focus on the estimation of four loss functions in GBT: mean square error (MSE), mean absolute error (MAE), Huber loss and quantile loss. They worked with a dataset of different known features and split this dataset into two groups: one for training and one for testing. Then they trained different GBT models with each loss function and evaluated their prediction accuracy on the test set.

Ho et al [14] employed three machine learning algorithms for predicting property prices: support vector machine (SVM), random forest (RF) and gradient boosting machine (GBM). They performed these models on a data set of 40,000 property transactions in Hong Kong for 18 years and compared the results of these models. The performance of their models was measured using three metrics: mean squared error (MSE), root mean square error (RMSE) and mean absolute percentage error (MAPE).

Zou [15] developed predictive model using a variety of methods including logistic regression, support vector regression, lasso regression, and decision trees. The study collects data from 3,000 properties in Jinan and uses R-squared to evaluate the effectiveness of these algorithms. The authors' study provides valuable insights into the use of machine learning to accurately predict house prices in the specific context of Jinan, China.

3. Methodology

The objective of this research is to develop an approach for predicting house prices using three machine learning algorithms, in this section we will detail the different methods used to carry out this project.

3.1. Linear regression

Linear regression is a widely utilized statistical model that enables the prediction of dependent variables based on the input of multiple independent variables [8]. By

employing least squares, linear regression establishes a linear equation. This equation effectively describes the relationship between the independent and dependent variables, serving as the foundation for constructing the prediction model.

Equation: $y = mx + b$

y is dependent variable.
x is independent variable.
m is estimated slop.
b is estimated intercept.

3.2. Random Forest Regression

Random forest regression, as the name implies, uses several decision trees to create an ensemble model that collectively generates predictions. The reason for using this type of regression is that because the trees are created in parallel and are relatively uncorrelated, good results can be obtained because each tree is not affected by the individual errors of the other trees. The algorithm (random forest) determines the results depending on the predictions of the decision trees. The algorithm uses the average of the results of several trees to make predictions. Increasing the number of trees allows to increase the accuracy of the results. Understanding this scheme, the random forest algorithm combines the results of several decision trees to produce a final result.

3.3. Gradiant Boosting Regression

Gradient boosting regression is a machine learning algorithm that combines iteratively weak prediction models (usually decision trees) to achieve accurate predictions for regression problems. It corrects the error of the previous model by training a new prediction model on the residuals of the existing set of predictions. Calculates the negative gradient of the loss function to determine the direction and magnitude of updates to the model parameters. The final prediction is obtained by the addition of the predictions of all the models in the ensemble. Gradient regression boosting is recognized for its high predictive performance and its ability to handle complex data sets.

3.4. Dataset

The dataset used is the Melbourne housing dataset disponible on Kaggle, which contains comprehensive information on residential properties in Melbourne, Australia. This dataset offers a variety of characteristics that are valuable for analysing and understanding the local housing market. These characteristics are as follows: Suburb, Address, Rooms, Type, Price, Method, Seller G, Date, Distance, Postcode, Bedroom2, Bathroom, Car, Land size, Building Area, Year Built, Council Area, Latitude, Longitude, Region name and Property count.

3.5. Data Analysis

Analyzing a dataset before building a model is an important step in better understanding the data and its components. As part of the analysis, a correlation matrix (see Figure 1) was plotted to examine the relationships between the various attributes. The correlation matrix contains correlation numbers from +1 to -1, indicating the degree of relationship between the two variables. A positive correlation indicates a positive relationship, while a negative correlation indicates an inverse relationship. A zero correlation, on the other hand, means that there is no relationship, and the variables are independent. Examining the correlation matrix allows us to better understand the interdependence between attributes and make informed decisions in the modeling process.

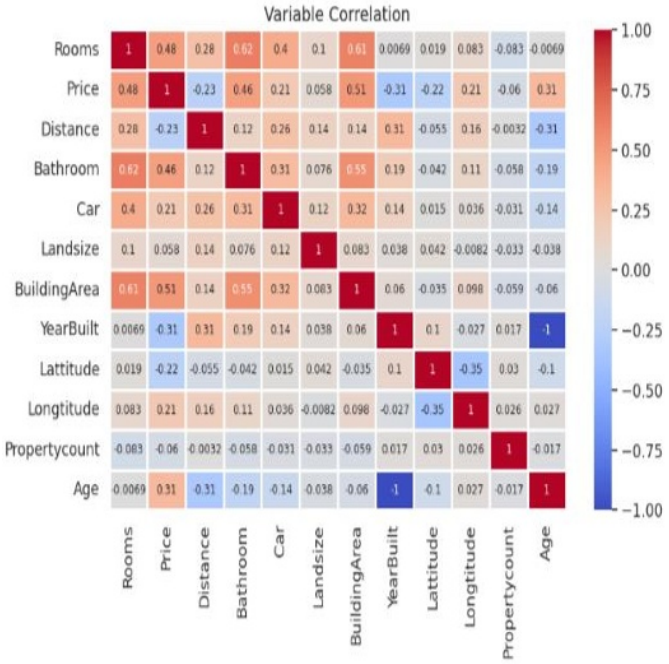

Fig.1. Correlation Matrix

3.6. Data Pre-Processing

We have applied some processes on the dataset before launching our models:

- Convert Object Columns to Categorical
- Remove Unnecessary Columns (Two columns, Bedroom2 and Rooms v Bedroom2, are dropped from the dataset. The 'Rooms v Bedroom2' column is created as a difference between the 'Rooms' and 'Bedroom2' columns, but it is determined to be unnecessary and removed)

- Add Age Variable (A new variable, 'Age', is added to the dataset by calculating the difference between the current year (2023) and the YearBuilt column)
- Identify Historic Homes (Based on the Age variable, homes with an age of 50 or more are classified as 'Historic', while others are labelled as modern)
- Convert Historic Column to Category
- Remove Rows with Missing Data
- Apply Standard Scaling to Numeric Variables
- Encode Categorical Variables
- Split the Data into training and testing sets.

3.7. Performance Metrics

We have used three performance metrics: mean absolute error Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2 Score), which are detailed in the following.

3.7.1. The mean absolute error (MAE):

is obtained by averaging the errors using absolute operations without distinguishing between plus and minus [16,17]. It calculates the difference in absolute value between the predicted and actual data and the test data average. The MAE equation is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \quad (1)$$

Where n representing the number of samples, y_i predicted value, and x_i the actual value.

3.7.2. Root Mean Squared Error (RMSE)

Root Mean square error (RMSE) is a widely used metric for evaluating the performance of regression models. It is a measure of the mean difference between the predicted values of the model and the actual values of the dataset. The RMSE equation is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}} \quad (2)$$

where:
n denotes the total number of observations.
x_i represents the actual values.
$\hat{x}_i$ represents the predicted or estimated values.

3.7.3. Coefficient of Determination (R^2 Score)

The coefficient of determination (R^2 Score) serves to determine the relationship between two input variables [12]. R2 can also be seen as the fractional difference between the model's prediction in the numerator and the mean minus 1 in the denominator. R2 ranges from 0 to 1; a score of 1 implies that the model is optimal, while a score of 0 implies that the model is suboptimal. The R^2 Score equation is as follows:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y})^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2} \quad (3)$$

where:

y_i denotes the observed values.
$\hat{y}$ represents the predicted or estimated values.
$\bar{y}$ denotes the mean value of the observed values.
The equation for the mean of all observed values $\bar{y}$ is as following:

$$\bar{y} = \frac{1}{n}\sum_{i=0}^{n-1} y_i \quad (4)$$

4. Results

The results shown in Table 1 show the performance of the three models, Linear Regression, Random Forest, and Gradient Boosting, was evaluated using three metrics: root mean square error, root mean square error (RMSE) and mean absolute error (MAE).
The Gradient Boosting model performed better than the others. Its estimated maximum R-square was 0.828, which means that it explains most of the variability in the housing dataset. In addition, the model has lower error values, RMSE and MAE. These lower error values indicate that the Gradient Boosting model provides a more predictive performance than other models. In general, the Gradient Boosting model outperforms the linear regression and Random Forest models in terms of R-Square and error values and is therefore the most appropriate model for predicting house prices in the Melbourne database.

Model	RMSE	R^2 Score	MAE
Linear Regression	372291.53505	0.639	272953.5787
Random Forest	261166.3075	0.822	171486.2709
GradientBoosting	257062.4425	0.828	169009.3626

Table 1. Results of different models on Melbourne housing dataset

Fig2. predicted values from LR Vs actual values.

Fig3. predicted values from RF Vs actual values.

Fig4. predicted values from GB Vs actual values.

Figs 2, **3** and **4** represent the predicted values of these three models as a function of the actual values. The Random Forest and Gradient Boosting models are very similar and show better prediction performance than the linear regression model.

5. Conclusion

In this research, three machine learning algorithms (Linear regression, Random Forest, and gradient boosting) are employed for predicting house prices using Melbourne housing dataset. We have described a step-by-step the process for analyzing the dataset and determining correlations between attributes before applying our models. Three metrics are used to evaluate the performance of models: Mean Absolute Error (MAE) Root Mean Squared Error (RMSE) Coefficient of Determination (R2 Score). we found that the gradient boosting gives better performance than the other models.in future work we will combine these three models for improving the performance.

Reference

[1] Alfiyatin, A. N., Febrita, R. E., Taufiq, H., & Mahmudy, W. F. (2017). Modeling house price prediction using regression analysis and particle swarm optimization case study: Malang, East Java, Indonesia. International Journal of Advanced Computer Science and Applications, 8(10).

[2] Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy, 111, 104919.

[3] Greenaway-McGrevy, R., & Sorensen, K. (2021). A Time-Varying Hedonic Approach to quantifying the effects of loss aversion on house prices. Economic Modelling, 99, 105491.

[4] Filip, F. G., Zamfirescu, C. B., & Ciurea, C. (2017). Computer-supported collaborative decision-making. Cham: Springer International Publishing.

[5] Kayode, A. A., Akande, N. O., Adegun, A. A., & Adebiyi, M. O. (2019). An automated mammogram classification system using modified support vector machine. Medical Devices: Evidence and Research, 275-284.

[6] Aderonke, K., Oluwatobi, A., Jabaru, S., & Tinuke, O. (2020). An Empirical Investigation of the Prevalence of Osteoarthritis in South West Nigeria: A Population-Based Study.

[7] Noah Akande, O., Christiana Abikoye, O., Anthonia Kayode, A., & Lamari, Y. (2020). Implementation of a

framework for healthy and diabetic retinopathy retinal image recognition. Scientifica, 2020.
[8] Tékouabou Koumétio, S. C., & Toulni, H. (2021). Improving knn model for direct marketing prediction in smart cities. In Machine Intelligence and Data Analytics for Sustainable Future Smart Cities (pp. 107-118). Cham: Springer International Publishing.
[9] Tékouabou, S. C., Gherghina, Ș. C., Toulni, H., Mata, P. N., & Martins, J. M. (2022). Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods. Mathematics, **10**(14), 2379.
[10] Shinde, N., & Gawande, K. (2018). Valuation of house prices using predictive techniques. Journal of Advances in Electronics Computer Science, 5(6), 34-40.
[11] Dagar, A., & Kapoor, S. (2020). A Comparative Study on House Price Prediction. International Journal for Modern Trends in Science and Technology, 6(12), 103-107.
[12] Jha, S. B., Pandey, V., Jha, R. K., & Babiceanu, R. F. (2020). Machine learning approaches to real estate market prediction problem: a case study. arXiv preprint arXiv:2008.09922.
[13] Hjort, A., Pensar, J., Scheel, I., & Sommervoll, D. E. (2022). House price prediction with gradient boosted trees under different loss functions. Journal of Property Research, 39(4), 338-364.
[14] Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 48-70.
[15] Zou, C. (2023). The House Price Prediction Using Machine Learning Algorithm: The Case of Jinan, China. Highlights in Science, Engineering and Technology, **39**, 327-333.
[16] Tékouabou, S. C. K., Chabbar, I., Toulni, H., Cherif, W., & Silkan, H. (2022). Optimizing the early glaucoma detection from visual fields by combining preprocessing techniques and ensemble classifier with selection strategies. Expert Systems with Applications, **189**, 115975.
[17] Glen, S. (2020, December 28). Absolute Error & Mean Absolute Error (MAE). Statistics How To. Retrieved from https://www.statisticshowto.com/absolute-error/