Journal of Contemporary Issues and Thought
ISSN 2232-0032/ e-ISSN 0128-0481/ Vol 13, Issue 1, 2023 (35-40)

The Prediction of Affordable Housing Prices in Petaling Jaya District Using R Statistical Computing Environment

Suresh Nodeson[1]*, Kesavan Krishnan[2], & Sathis Krishnan[3]

[1]*Faculty of Business and Finance, Universiti Tunku Abdul Rahman, Jalan Universiti, Bandar Barat, 31900 Kampar, Perak, Malaysia.*
[2]*Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Jalan Universiti, Bandar Barat, 31900 Kampar, Perak, Malaysia.*
[3]*Faculty of Computer Science and Information Technology, University of Malaya, 50603 Lembah Pantai, Kuala Lumpur, Malaysia.*
**Email: suresh@utar.edu.my*

DOI: https://doi.org/10.37134/jcit.vol13.1.4.2023

To cite this article (APA): Nodeson, S., Krishnan, K., & Krishnan, S. (2023). The Prediction of Affordable Housing Prices in Petaling Jaya District Using R Statistical Computing Environment. *Journal of Contemporary Issues and Thought*, *13*(1), 35–40. https://doi.org/10.37134/jcit.vol13.1.4.2023

Abstract

Affordable housing especially in a city environment is recognized as one of citizen needs among the mid-dle-income groups. This research paper intended to explore the possibilities to use data mining algorithms: Linear Regression, Random Forest and Gradient Boosting algorithms for predicting and analyzing the housing affordability price for middle-income earners in Petaling Jaya district, Malaysia. The dataset from Malaysia House Index by Petaling Jaya district used to evaluate based on the proposed algorithms. The dataset extracted based on residential property sub-sectors with three attributes. Based on the prediction models, as results, the Gradient Boosting algorithm shows higher accuracy of 74% for predicting affordable housing price in Petaling Jaya district, Malaysia compared to other prediction techniques.

Keywords: Affordable housing price; Linear regression; Random forest and gradient boosting

1. Introduction

Everyone wants to be able to acquire a home at a reasonable price, especially if they live in a developed region. The price of property always goes up or down unexpectedly, and most economists cannot foresee these price fluctuations in detail. This change does not exclude middle-income consumers. However, many individuals still have no experience estimating property prices, and the desire to buy a property in this location remains a desire (Chan, 2001). This is a significant issue since they cannot estimate the price of a property in a certain region (Zietz et al., 2008). As a result, the goal of this study is to develop and compare data mining algorithms for forecasting affordable home prices in the Pearling Jaya district.

Property price alone cannot determine any property price; it must also be influenced by other indirect elements, including the present state of the economy, politics, and some social influences (Wood and Stockhammer, 2020). In addition to the direct elements, there are certain indirect elements that might impact the price of a property, such as the distance to the nearest public transportation, the location of the home and its surroundings, and property kinds (Stokenberga, 2014). As a result, in order to anticipate any property price,

these direct elements need also be incorporated in the prediction process.

There have been several algorithms used to anticipate affordable housing price, such as the neural network method, which is used to predict properties in Johor, although the author ignored the model's accuracy (Zainun et al., 2010). In order to have an appropriate sales price for a property, many algorithms must be implemented. Therefore, the study implemented three separate sets of algorithms, RIPPER, Naive Bayesian, and Ada-Boost, to forecast the sale price of the property, with the RIPPER algorithm showing a higher accuracy rate than the other two (Jaiswal and Patil, 2020; Wipf and Rao, 2007). The authors of the other research, on the other hand, utilized hedonic regression and artificial neural networks to estimate home prices, but the findings reveal that the artificial neural network algorithms surpass the hedonic algorithm in terms of accuracy (Selim, 2009).

Based on the finding above, we can infer that the vast majority of research implemented various types of algorithms to forecast affordable home prices. However, these algorithms are evaluated using the same size testing set rather than separate sets of testing sets. Therefore, the purpose of this study is to select the Linear Regression, Random Forest, and Gradient Boosting algorithms to forecast affordable housing prices and their performance.

1.1 Dataset

The property sales data from National Property Information Centre, Malaysia were used to perform prediction analysis. The data source has been extracted based on historical data with the last five years (Q1,2016 – Q1,2021) for Petaling Jaya district with the following details of datasets:

Table 1: List of housing price datasets for Petaling Jaya district

No.	Dataset Name	Source	Attributes Types	Number of Attributes
1	Number and Percentage of Transactions by Price Range for The Principal Property Sub-Sectors	National Property Information Centre (valuation and property services department)	Discrete and continuous	3
2	Breakdown Value of Residential Property Transactions According to Type, Price Range and District (RM MILLION)	National Property Information Centre (valuation and property services department)	Discrete and continuous	7

The following Table 2 shows the list of attributes that extracted from the datasets:

Table 2: List of attributes with description

Dataset Name	Attributes No.	Description
Number and Percentage of Transactions by Price Range for The Principal Property Sub-Sectors	1	Years
	2	Price range
	3	Residential number
Breakdown Value of Residential Property Transactions According to Type, Price Range and District (RM MILLION)	1	Years
	2	Property type
	3	Price Range
	4	District
	5	Distance to The Nearest Public Transportation
	6	Location of The Home and Its Surrounding

2. Proposed Method

This section explains the suggested workflow that is involved in this statical analytical process, which is built using an input process output (IPO) architecture. The steps in this procedure are depicted in Figure 1, which begins with the extraction of a housing raw dataset and finishes with the evaluation of the housing price.

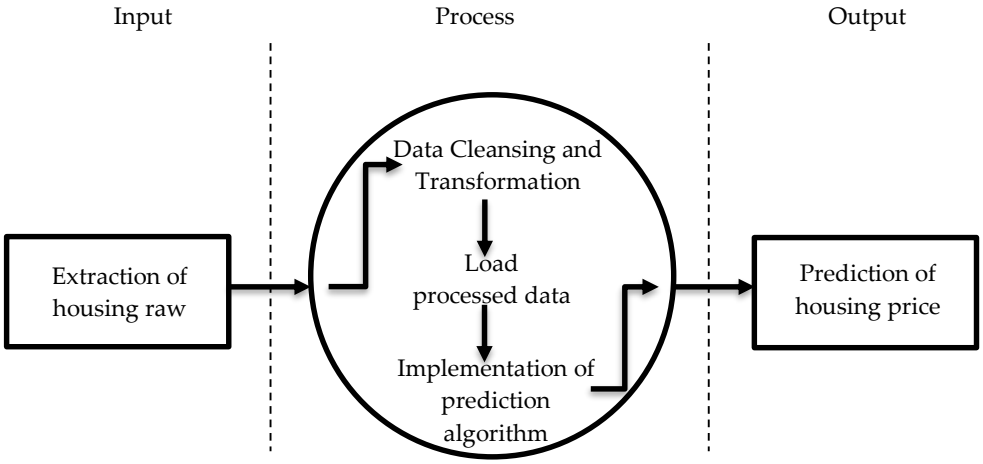

Figure 1: The steps involved in the proposed affordable housing price forecast

2.1 Data Cleansing and Transformation (Pre-Process of Dataset)

The process of eliminating irrelevant data from a raw dataset and transforming it into meaningful information in an appropriate format in order to enhance the dataset's consistency and correctness is known as data cleaning and transformation. This procedure necessitates a number of cleaning steps, including the removal of duplicate entries, the elimination of unnecessary data, the correction of structural problems, and the management of missing data. This has the potential to deliver benefits by retaining data quality in terms of consistency and accuracy.

2.2 Training and Testing Dataset

The affordable housing pricing dataset has been divided into two different sets: the training dataset (80%), and the testing dataset (20%). The testing dataset is divided into three different sizes as well: 10%, 20%, and 30%. The following Figure 2 depicts the dataset splitting procedure.

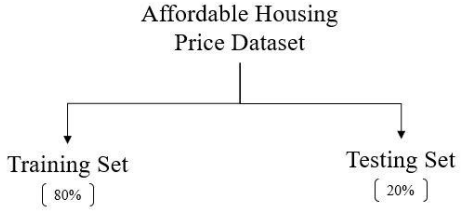

Figure 2: The dataset splitting procedure: Training Set and Testing Set

2.3 Types of Data Mining Algorithms used for Affordable Housing Price Prediction

This study employs three distinct sets of data mining algorithms, which are explained below:

(A) Linear Regression: This algorithm is used to estimate the relationship between several variables, and this relationship is calculated based on current data to provide statistical relationships. In terms of their relationship, the dependent and independent variables might be constructed appropriately (Ghosalkar and Dhage, 2018).

(B) Random Forest: The random forest algorithm, also known as regression forest, is used to predict both regression and classification, and its basic function is to create decision trees based on a random selection of input values from the dataset, resulting in a decision forest. The decision forest findings were merged to provide the present estimates. Apart from their capacity to provide credible predictions, the key justification for choosing these algorithms for this study is their use in previous similar studies (Hong et al., 2020).

(C) Gradient Boosting: Like the random forest, this algorithm is used for both regression and classification, and it essentially merge many types of simple regression models into a composite single model (Sangani et al., 2017). Apart from that, this algorithm is composed of three independent sets of elements: a loss function, a weak learner, and, ultimately, an additive model. The loss function may be defined by concentrating on the goal of the issue being addressed, and we may estimate it using Mean Square Error (MSE) and Mean Absolute Error (MAE). Meanwhile, weak learners might concentrate on the decision trees. Because of the algorithm's benefit in terms of procedures that can be repeated, the basic regression predictor of the dataset can be learnt and then the residual error can be estimated (Taieb and Hyndman, 2014).

3. Results

This section presents the results based on the three model performances: linear regression, random forest regressor, and gradient boosting. Different sizes of testing sets were used to forecast the accuracy rate of each model for the property data set using the holdout approach, as shown in Table 3 with the accuracy rate of each model with the sizes. When compared to the other algorithms, the GPR method has a better accuracy rate (74 percent) for the 20% size of the testing set. This demonstrates that the size of the testing set is accepted as 20% for these prediction models.

Table 3: Overall accuracy rate results for prediction of affordable housing price

Types of Algorithms	Accuracy rate of affordable housing price		
	10%	20%	30%
Linear Regression	0.59	0.63	0.54
Random Forest Regressor	0.66	0.71	0.68
Gradient Boosting	0.72	0.74	0.70

Despite the fact that the testing set size is 30%, the linear regression algorithms have the lowest accuracy rate (52%) when compared to the other two methods. Meanwhile, when the size of the testing set is reduced to 20%, the linear regression models attain an accuracy rate of approximately 64%. Table 4 displays the overall performance of all prediction models based on the following metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Square Error (MSE) (MSE).

Table 4: Performance comparison for prediction of affordable housing price

Types of Algorithms	MAE	MSE	RMSE
Linear Regression	5.26	51.20	7.24
Random Forest Regressor	4.81	38.62	6.61
Gradient Boosting	3.94	35.07	5.17

According to the values of MAE, RMSE, and MSE, gradient boosting has the lowest value when compared to linear regression, indicating that the gradient boosting algorithms forecast property prices in the Petaling Jaya district more correctly than the other two models.

4. Conclusion

In this study, three distinct data mining algorithms were used to forecast the pricing of affordable housing: linear regression, random forest, and gradient boosting. This prediction was made using the training set (80% of the dataset) and the testing set (20% of dataset). We observed that the gradient boosting algorithms surpasses the other two algorithms in terms of house pricing rate in the Petaling Jaya district based on these forecasts. As a result, the gradient boosting algorithms may be improved and expanded in the future to anticipate reasonable home costs.

References

Chan, S. (2001). Spatial lock-in: Do falling house prices constrain residential mobility? *Journal of Urban Economics*, *49*(3), 567-586.

Ghosalkar, N. N., & Dhage, S. N. (2018). *Real estate value prediction using linear regression*. Paper presented at the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).

Hong, J., Choi, H., & Kim, W. S. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management*, *24*(3), 140-152.

Jaiswal, K. B., & Patil, H. (2020). The study using ensemble learning for recommending better future investments. *International Journal of Advanced Research in Computer Science*, *11*(6), 23-32.

Sangani, D., Erickson, K., & Al Hasan, M. (2017). *Predicting zillow estimation error using linear regression and gradient boosting*. Paper presented at the 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS).

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*(2), 2843-2852.

Stokenberga, A. (2014). Does bus rapid transit influence urban land development and property values: A review of the literature. *Transport Reviews*, *34*(3), 276-296.

Taieb, S. B., & Hyndman, R. J. (2014). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting*, *30*(2), 382-394.

Wipf, D. P., & Rao, B. D. (2007). An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Transitions on Signal Processing*, *55*(7), 3704-3716.

Wood, J., & Stockhammer, E. (2020). House prices, private debt and the macroeconomics of comparative political economy. Working Papers PKWP2005, Post Keynesian Economics Society (PKES).

Zainun, N. Y. B., Rahman, I. A., & Eftekhari, M. (2010). Forecasting low-cost housing demand in Johor Bahru, Malaysia using artificial neural networks (ANN). *Journal of Mathematics Research*, 2(1), 14-19.

Zietz, J., Zietz, E. N., & Sirmans, G. S. (2008). Determinants of house prices: A quantile regression approach. *Journal of Real Estate Finance and Economics*, *37*(4), 317-333.