p-ISSN 2338 – 3216
e-ISSN 2528 – 1070

JURNAL STATISTIKA
UNIVERSITAS MUHAMMADIYAH SEMARANG
12(2) 2024: 01-11

DECISION TREE-BASED GRADIENT BOOSTING: ALGORITHM TO APPROACH HOUSE PRICE PREDICTION IN JAKARTA, BOGOR, DEPOK, TANGERANG, AND BEKASI (JABODETABEK)

Intan Lisnawati[1*], Anjasmoro Adi Nugroho[2]
[1] Mathematics, Faculty of Science, National Central University, Taiwan
[2] Software Engineering, Faculty of Software Engineering, Innopolis University, Russia

***e-mail**: *intanlisnawati17@gmail.com*

Article Info:
Received: September 24th, 2024
Accepted: October 24th, 2024
Available Online: December 15th, 2024

Keywords:
Boosting, ensemble method, house price, prediction, root mean square error
.

Abstract: The house sale prices are a particular concern for some people, whether sellers or buyers, for personal use or investment. Commonly, the buyer comes from newly-married couples, parents, or investors. Compared to years ago, the recent price is more expensive due to some conditions over the time. Forecasting is a method to see at which price the house may fit the market price with certain features. Through this study, we complement the previous research about house prices and analyze the results. Besides, here we also break down the algorithm and sketch the steps so that it eases the reader to understand the method. Exploratory data analysis is also done to see and analyze the characteristics of the dataset. Applying decision tree-based gradient boosting, we run the algorithm into datasets in Jakarta, Bogor, Depok, Tangerang, and Bekasi (Jabodetabek) consisting of house price and its features. We see that the RMSE value is Rp277.369.397 and the MAPE is 17,37%. With that value of accuracy we could mention that gradient boosting is quite competitive compared with other methods and able to give its best prediction over house prices.

1. INTRODUCTION

Jakarta, Bogor, Depok, Tangerang, and Bekasi or abbreviated as Jabodetabek is well-known as an area which has many companies. In Jakarta itself, thousands of companies are there, ranging from food companies to pharmaceutical companies [1]. Besides, Jabodetabek is an area that has a minimum wage with a nominal value that is relatively high compared to other areas in Indonesia. Looking at the Ministry of Manpower of Republik Indonesia website, in 2024 Jakarta got the highest provincial minimum wage, that is Rp5.067.381 [2]. Meanwhile, the minimum wage in Bogor is Rp4.813.988, Depok is Rp4.878.612, Tangerang is Rp4.760.289, and Bekasi Rp5.343.430, [3]. These 5 areas are included in the top 8 category with the highest regional minimum wage in Indonesia. The things mentioned above are probably what drives Jabodetabek to be a densely populated area in Indonesia. The high population in Jabodetabek, encourages the availability of housing there as a place to live.

Housing is categorized as a primary need in life [4]. There are many aspects that influenced the house price, e.g. location, building area, total bedrooms and bathrooms, carport, etc. Several researchers worked using varied algorithms, such as random forest [5], general

[DOI: 10.14710/JSUNIMUS.12.2.2024.1-11]

regression neural network [6], linear regression [7], [8], [9], and multiple linear regression [10] had been conducted to see how algorithms work giving a house sale price prediction.

Recently, machine learning has become a popular topic for students. A lot of machine learning methods were used by researchers to see the effectiveness of the algorithms. One of the most commonly used methods in machine learning is gradient boosting, as it considers previous steps to give an estimation. Boosting itself is categorized as an ensemble learning method, as it also has another type called bagging [11].

Ensemble learning is a method which combines a set of learners, so that the predictor gives the smallest loss. It involves a combination of techniques that allows multiple machine learning models, called base learners (or, sometimes, weak learners), to consolidate their predictions and output a single, optimal prediction, given their respective inputs and outputs [12]. Also, it builds the model in a sequential manner such that the new model is built to correct the errors made by its predecessor [11].

In boosting, the base learners can be any machine learning algorithm that performs slightly better than random guessing, including decision trees, linear regression models, or support vector machines [13]. Gareth also mentioned that boosting is able to use decision trees as its base learner [14]. Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [15]. Shwartz and David [16] have argued that supervised learning describes a scenario in which a training example contains significant information that is missing in the unseen "test examples" to which the learned expertise is to be applied. Ghevira [17] applied decision tree and gradient boosting to a regression case due to predicting life expectancy. In this study, we consider decision trees to be the base learners.

Knowing the advantage of ensemble learning especially in gradient boosting, we are interested to complement the previous research in the field of gradient boosting works to give house price prediction and investigate the precision for the prediction with the real price especially in Jabodetabek with parameters included on the dataset. The model is first trained by using a training data set until it reaches the desired model, through the smallest loss between observed and predicted value. Also, since so far we have not seen the complete steps written in mathematics as the way to find the prediction, thus in this study the algorithm was broken down to be written as steps, so that it eases the reader to understand the method.

2. LITERATURE REVIEW

2.1. Gradient Boosting

Gradient boosting was invented by an American statistician, Jerome Harold Friedman in 1999 through his paper, "Greedy Function Approximation: A Gradient Boosting Machine" [18]. In this section we provide a brief and straightforward overview of theories and statements related to gradient boosting algorithms which are divided into 2 parts as the following subsection.

2.1.1 Gradient Points in The Direction of Maximum Increase

Definition 1 *Let $z = f(x, y)$ be a function of x and y such that f_x and f_y exist. The vector $\nabla f(x, y)$ is called the* ***gradient*** *of f and is defined as*

[DOI: 10.14710/JSUNIMUS. 12.2.2024.1-11]

$$\nabla f(x,y) = f_x(x,y)i + f_y(x,y)j \tag{1}$$

The vector $\nabla f(x,y)$ *is also written as "grad. f" [19].*

Theorem 1 Properties of the gradient, see [19].

i. If $\nabla f(x_0,y_0) = 0$*, then* $D_u f(x_0,y_0) = 0$ *for any unit vector u.*

ii. If $\nabla f(x_0,y_0) \neq 0$*, then* $D_u f(x_0,y_0)$ *is maximized when u points in the same direction as* $\nabla f(x_0,y_0)$*. The maximum value of* $D_u f(x_0,y_0)$ *is* $||\nabla f(x_0,y_0)||$.

iii. If $\nabla f(x_0,y_0) \neq 0$*, then* $D_u f(x_0,y_0)$ *is minimized when u points in the same direction as* $\nabla f(x_0,y_0)$*. The maximum value of* $D_u f(x_0,y_0)$ *is* $-||\nabla f(x_0,y_0)||$.

Since the gradient points in the direction of maximum increase, so to find the maximum decrease we will multiply g_m by -1 becomes $-g_m$ so that it will point in the opposite direction of steepest ascent or we can call it steepest descent. Therefore, $-g_m$ will help us to show in which direction we should move to get the smallest $\Phi(P)$. The current gradient $g_m$defined below will point at which direction we should move.

$$g_m = \left(\left[\frac{\partial \phi(P)}{\partial P_J} \right]_{P={P^*}_{m-1}} \right), \qquad 1 \leq j \leq v$$

$$= \left(\left[\frac{\partial \phi(P)}{\partial P_1} \right]_{P={P^*}_{m-1}}, \dots, \left[\frac{\partial \phi(P)}{\partial P_v} \right]_{P={P^*}_{m-1}} \right) \tag{2}$$

where

$${P^*}_{m-1} = \sum_{i=0}^{m-1} c_i \tag{3}$$

Meanwhile, how far the step we take along that $-g_m$direction is determined by the ρ_m value

$$\rho_m = arg \min_{\rho} \Phi({P^*}_{m-1} - \rho g_m) \tag{4}$$

A brief simulation to see how gradient works is shown on the figures below with setting $v = 2$ and $M = 2$.

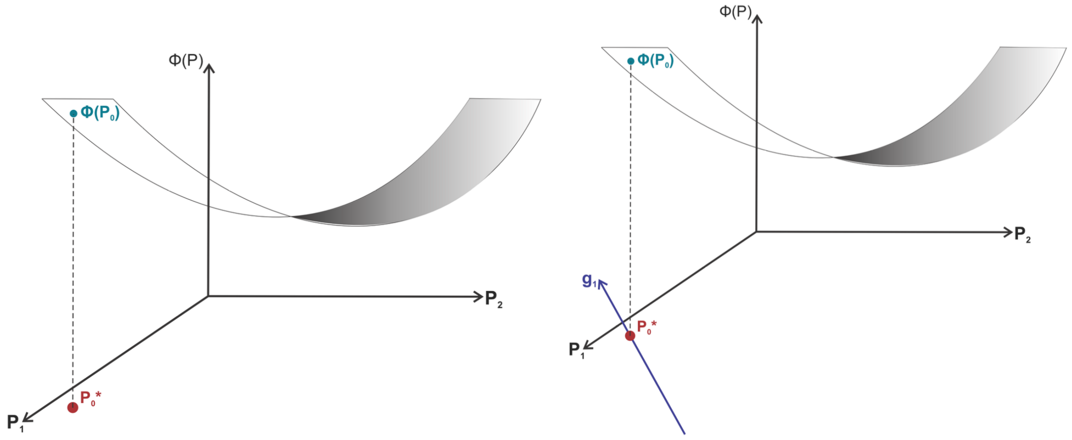

Fig 1. Initial Step of Gradient Boosting and Its Next Gradient Candidate g_1

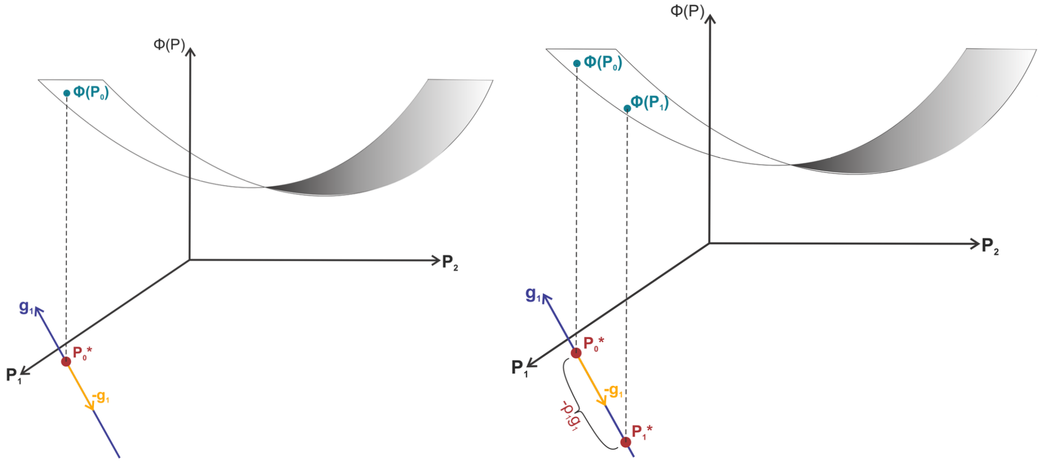

Fig 2. The Negative Gradient $-g_1$ Heading To The Opposite Direction of g_1 along ρ_1

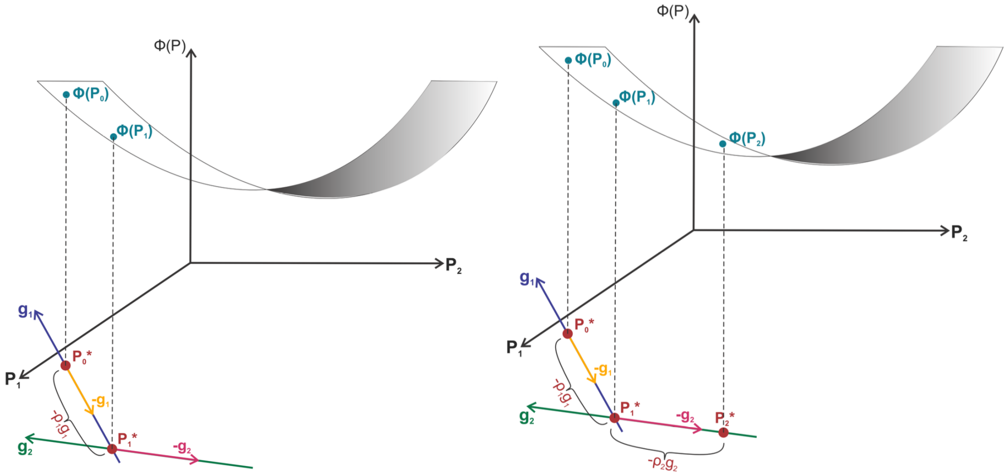

Fig 3. The Next Gradient Candidate g_2 and Its Opposite Direction along ρ_2

2.1.2 Algorithm Steps

If we sort the steps of gradient boosting algorithm, it will give us the following sequential steps [20]:

1. Given data set S: $\{(x_i, y_i);\ x_i \in \Re^k, y_i \in \Re, 1 \leq i \leq N\}$.
2. Set an initial prediction,

$$F_0(x), \text{ with } F_0(x) = \underline{y} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{5}$$

3. Calculate

$$-g_1(x_i) = -\left[\frac{\partial l(y,z)}{\partial z}\right]_{y=y_i, z=F_0(x_i)} \tag{6}$$

with $l(y, z)$ is any differentiable loss function, and

$$\rho_1 \ = arg \min_{\rho} \sum_{i=1}^{N} l\left[y_{i,} F_0(x_i) + \rho h(x_i; \alpha_1)\right] \tag{7}$$

and

$$\alpha_1 = arg \min_{\alpha,\beta} \sum_{i=1}^{N} [-g_1(x_i) - \beta h(x_i; \alpha)]^2 \tag{8}$$

4. Then,

$$F_1(x) = F_0(x) + \rho_1 h(x; \alpha_1) \tag{9}$$

5. The following prediction will approach,

$$F_2(x) = F_1(x) + \rho_2 h(x; \alpha_2)$$

$$F_3(x) = F_2(x) + \rho_3 h(x; \alpha_3)$$

.

.

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; \alpha_m) \tag{10}$$

6. Repeat the process for each $m = 1, \ldots, M$ iteration.

2.2. Performance Measures

Performance measures are applied to see the accuracy of the method used to analyze the data. Refer to [19], here we use the Root Mean Square Error (RMSE) calculation to compare the accuracy between forecasting models and Mean Absolute Percentage Error (MAPE) to determine whether the forecasting results are accurate or not.

$$RMSE = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N} y_i - \underline{y}\right)^2} \tag{11}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \underline{y}}{y_i}\right| \tag{12}$$

[DOI: 10.14710/JSUNIMUS. 12.2.2024.1-11]

Smaller RMSE score indicates that the model is better and prefers to choose that model. Damaliana et al. [21] mentioned in their paper that a percentage of MAPE less than 10% indicates the forecast results are very accurate, 10% to 20% of the prediction results are accurate, 21% to 50% means reasonable forecast, and above 50% means the prediction results are not accurate.

3. METHODOLOGY

In this article, we apply the housing price dataset collected from *Kaggle*, see [22], an online platform which has a lot of dataset from various topics. The data was taken around the end of year 2022 with a total of 3.553 observations. Houses are sold in ranges starting from Rp42.000.000 until Rp580.000.000.000 with Rp4.191.685.000 on average. After preprocessing the data, the house price ranges from Rp400.000.000 until Rp3.000.000.000. The data distribution we used to analyze is represented in Figure 4. It can be seen that the data distribution is labeled as skewed-right histogram.

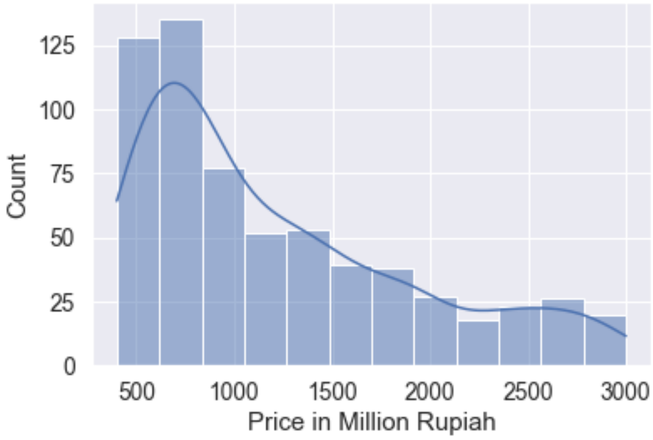

Fig 4. The Data Distribution of Jabodetabek's House Price

Completed with 25 features as mentioned in Table 1 below, these will result in the correlation of each feature to the house price.

Table 1. The House Features

Features				
District	Property Type	Building Size (m^2)	Maid Bedrooms	Building Orientation
Address	Facilities	Land Size (m^2)	Property Condition	Garages
City	Bedrooms	Certificate	Maid Bathrooms	Furnishing
Latitude	Bathrooms	Electricity	Building Age	Title
Length	Carports	Floors	Year Built	Ads ID

The numerical simulations and data analysis were run in *Python* program language. The original data was splitted into 90% training and 10% testing data set. To analyze the characteristics of the dataset, we also went through the data preprocessing which contains 'Exploratory Data Analysis'.

By using training data set, we sequentially train the data in $m = 1, \ldots, M$ iteration to get a model which gives us the smallest loss. Consequently, we run that model to predict the output of the information given in the testing data set. The accuracy between forecasting of each iteration is based on the RMSE value. Therefore, the smaller RMSE then the better the prediction is given by. The following chart will clarify the flow of the methodology in this research.

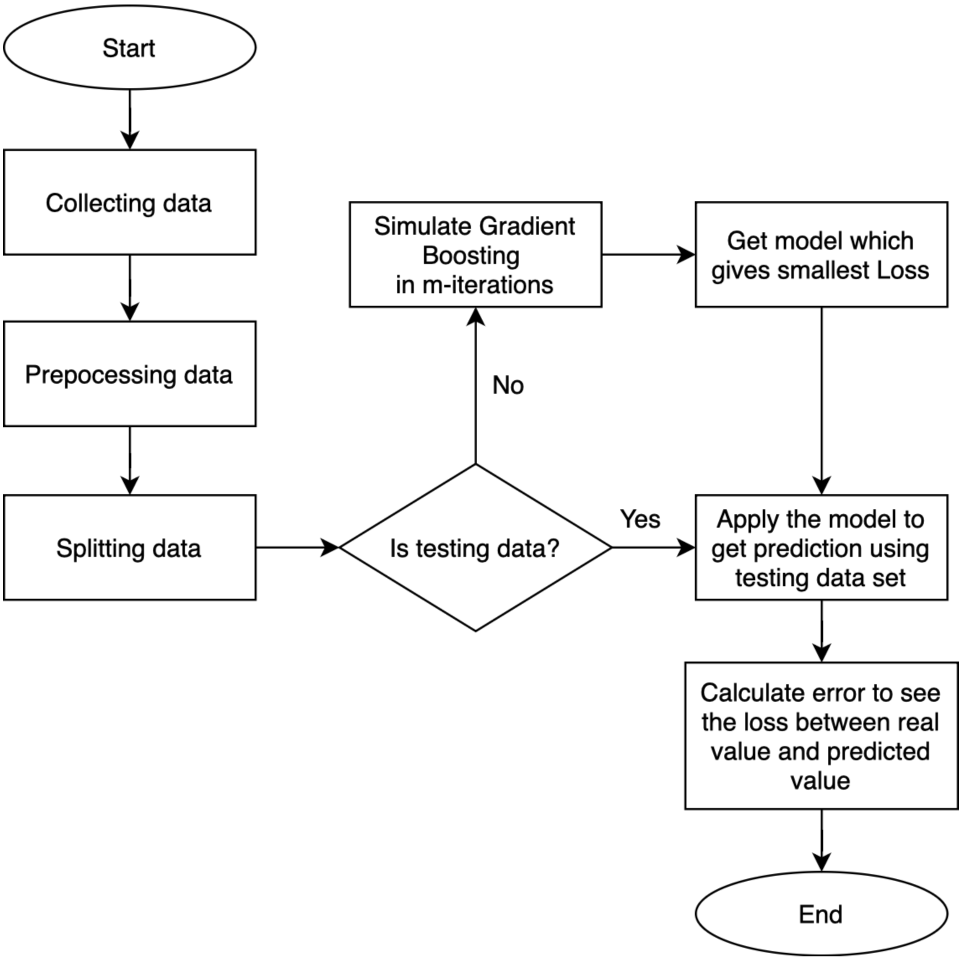
Start
Collecting data
Prepocessing data
Splitting data
Is testing data?
No
Yes
Simulate Gradient Boosting in m-iterations
Get model which gives smallest Loss
Apply the model to get prediction using testing data set
Calculate error to see the loss between real value and predicted value
End

Fig 5. Research Methodology Flowchart

4. RESULTS AND DISCUSSION

Since the gradient boosting method does an iteration step and gives an updated prediction value every time we iterate, thus we will run in some m iteration to approach the outcome. As can be seen in the iteration plot below, Figure 6, early simulations give higher values of RMSE meaning that the distance between prediction and real value is still too far. It keeps producing lower values as the iteration is more frequent which means that the distance of prediction is getting closer to the real value. The first iteration gives the largest RMSE, Rp643.924.702, meanwhile at 100^{th} iteration gives the lowest RMSE, Rp277.369.397. Calculating the MAPE at 100^{th} iteration, it results 17,37%. As a result, we took the model for prediction is at $M = 100$.

[DOI: 10.14710/JSUNIMUS. 12.2.2024.1-11]

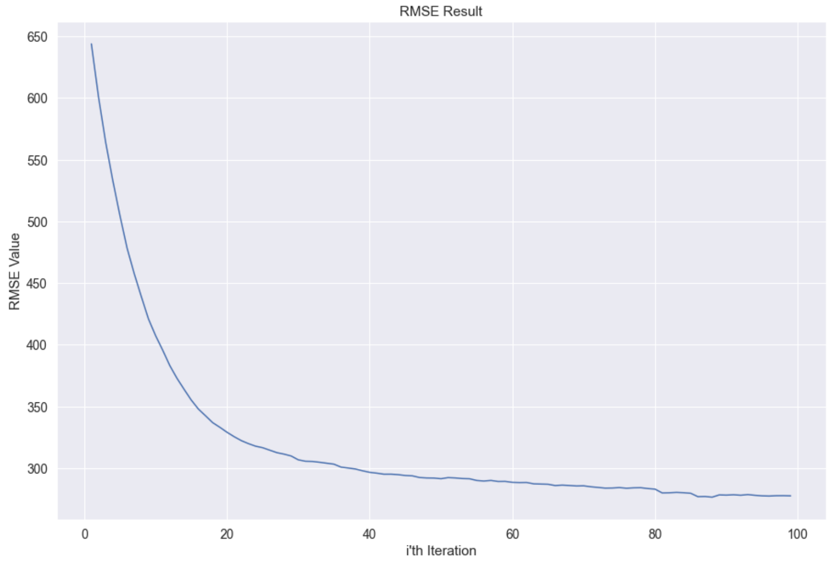

Fig 6. RMSE Value at Each Iteration

Figure 7 shows the plot of comparison between those values at 100^{th} iteration on the testing data. The prediction represented by the blue line is always almost in line to the red line which represents the real value. It means that our prediction has relatively small loss to the observed data. As a note, the tree we run here has a maximum depth of 3 as it is the default from the Scikit learn in *Python*.

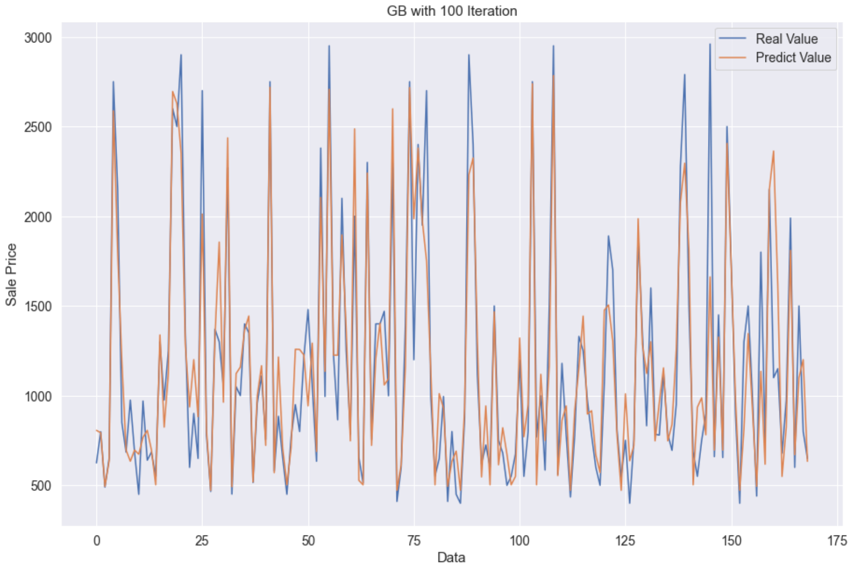

Fig 7. Comparison Plot Between The Original House Price and The Prediction Price

The MAPE score gives 17,37% or 82,63% accuracy. To see the gradient boosting performance, we compare to other methods in machine learning which also predict the same data set of Jabodetabek's house price. The result can be seen in Table 2 below which displays the algorithm, RMSE, MAPE, and accuracy.

Table 2. Comparison To Other Methods

Algorithm	RMSE	MAPE	Accuracy
Multi-Layer Perceptrons	Rp404.707.323	26.39%	73,61%
Linear Regression	Rp375.044.001	21.81%	78,19%
Decision Tree	Rp377.347.975	19.64%	80,36%
K-Nearest Neighbor (KNN)	Rp358.896.443	19.36%	80,64%
Gradient Boosting	Rp277.369.397	17,37%	82,63%
XGBoost	Rp275.451.547	16,80%	83,2%

If we see the above values, located at the second top, gradient boosting is able to give a better prediction than multi layer perceptrons, linear regression, decisionx tree, and K-Nearest Neighbor. It gives a prediction very close to the XGBoost. Meaning that, gradient boosting is quite competitive compared to other methods.

Since the base learner we apply in gradient boosting is the decision trees, we also provide the tree's appearance of the algorithm as can be seen in Figure 8. Here only shows a partial tree to illustrate the tree's construction due to the available space if it showed the full size trees clearly. The tree in gradient boosting keeps updating its prediction to fix the previous mistakes. It will stop building the tree following the initial set up.

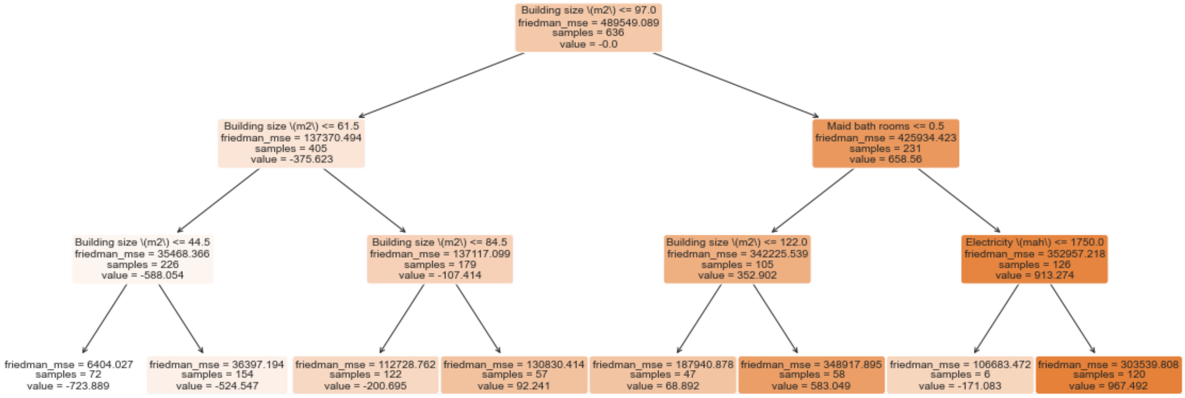

Fig 8. Partial Tree's Appearance in Gradient Boosting

5. CONCLUSION

Based on the RMSE result, the prediction of house price in Jabodetabek using gradient boosting is Rp277.369.397. Meanwhile, the accuracy of gradient boosting used to forecast subject to the MAPE is 17,37%. Even though the RMSE looks quite large, if we refer to [21], this score is still considered an accurate value in providing predictions. The author suspects that the large RMSE value is likely due to the margins in the property sector which are quite large and vary from one seller to another. People, especially those who live in Jabodetabek area, are still able to consider the method to see the house price with the certain features owned by each house. Therefore, it will help people to find residence based on their needs and budget. On the other side, sellers can set up an appropriate price following the characteristics of the house they sell.

This study limits only on the house price in Jabodetabek area due to the availability of the dataset. Any further study is welcomed, whether to explore using other methods, more parameters, or other areas so that the house price may vary and complement existing research.

REFERENCES

[1] Badan Pusat Statistik Provinsi Dki Jakarta, “Jumlah Perusahaan, Tenaga Kerja dan Pengeluaran Untuk Tenaga Kerja Menurut Klasifikasi Industri pada Industri Besar dan Sedang di Provinsi DKI Jakarta.” Accessed: Oct. 01, 2024. [Online]. Available: https://jakarta.bps.go.id/id/statistics-table/2/MzU5IzI=/jumlah-perusahaan--tenaga-kerja-dan-pengeluaran-untuk-tenaga-kerja--menurut-klasifikasi-industri-pada-industri-besar-dan-sedang-di-provinsi-dki-jakarta.html

[2] Kementerian Ketenagakerjaan RI, “Upah Minimum Provinsi Tahun 2024.” Accessed: Oct. 01, 2024. [Online]. Available: https://satudata.kemnaker.go.id/infografik/57

[3] Jobstreet Content Team, “12 Daerah dengan UMR Tertinggi di Indonesia Tahun 2024 - Jobstreet Indonesia.” Accessed: Oct. 01, 2024. [Online]. Available: https://id.jobstreet.com/id/career-advice/article/umr-tertinggi-di-indonesia#

[4] Friska Artaria Sitanggang and Prayetno Agustinus Sitanggang, *Buku Ajar Perilaku Konsumen*. Jawa Tengah: PT. Nasya Expanding Management, 2021.

[5] N. Hadi and J. Benedict, “Implementasi Machine Learning untuk Prediksi Harga Rumah Menggunakan Algoritma Random Forest,” *Computation: Journal of Computer Science and Information Systems*, vol. 8, no. 1, pp. 50–61, Apr. 2024.

[6] E. Febrion Rahayuningtyas, F. Novia Rahayu, and Y. Azhar, “Prediksi Harga Rumah Menggunakan General Regression Neural Network,” *Jurnal Informatika*, vol. 8, no. 1, pp. 59–66, Apr. 2021.

[7] A. Vermaysha and Nurmalitasari, “Prediksi Harga Rumah di Kabupaten Karanganyar Menggunakan Metode Regresi Linear Sistem Informasi,” *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB)*, pp. 6–11, Jul. 2023.

[8] A. Saiful, S. Andryana, and A. Gunaryati, “Prediksi Harga Rumah Menggunakan Web Scrapping Dan Machine Learning Dengan Algoritma Linear Regression,” *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, pp. 41–50, Mar. 2021.

[9] N. Nuris, “Analisis Prediksi Harga Rumah Pada Machine Learning Metode Regresi Linear Analisis Prediksi Harga Rumah Pada Machine Learning Menggunakan Metode Regresi Linear,” *Jurnal Explore*, vol. 14, pp. 108–22, Jul. 2024.

[10] M. Labib Mu’tashim, S. A. Damayanti, H. N. Zaki, T. Muhayat, and R. Wirawan, “Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression,” *Jurnal Informatik*, vol. 17, pp. 238–245, Dec. 2021.

[11] H. Wu, J.-M. Yamal, A. Yaseen, and V. Maroufy, *Statistics and Machine Learning Methods for EHR Data*. Chapman and Hall/CRC, 2020.

[12] George. Kyriakides and K. G. . Margaritis, *Hands-on ensemble learning with Python : build highly optimized ensemble machine learning models using scikit-learn and Keras*. Birmingham: Packt Publishing, 2019.

[13] D. Kharkar, “About boosting and gradient boosting algorithm.” Accessed: Oct. 11, 2024. [Online]. Available: https://www.linkedin.com/pulse/boosting-gradient-algorithm-dishant-kharkar/

[14] G. James, D. Witten, T. Hastie, and R. Tibshirani, “An Introduction to Statistical Learning with Applications in R Second Edition,” 2021.

[15] “1.10. Decision Trees.” Accessed: Oct. 10, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/tree.html

[16] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press, 2014. [Online]. Available: http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

[17] G. Chairunisa *et al.*, "Life Expectancy Prediction Using Decision Tree, Random Forest, Gradient Boosting, and XGBoost Regressions," *Jurnal Sintak*, vol. 2, no. 2, pp. 71–82, 2024.

[18] M. Chaturvedi, "Story of Gradient Boosting: How It Evolved Over Years," Analytics India Mag. Accessed: Oct 10, 2024. [Online]. Available: https://analyticsindiamag.com/ai-origins-evolution/story-of-gradient-boosting-how-it-evolved-over-years/

[19] G. Strang and E. (Jed) Herman, "Directional Derivatives and the Gradient - Calculus Volume3," Openstax. Accessed: Oct. 10, 2024. [Online]. Available: https://openstax.org/books/calculus-volume-3/pages/4-6-directional-derivatives-and-the-gradient

[20] I. Lisnawati, "Tree-based ensemble methods with an application in house sale price prediction," M.S. thesis, Dept. of Math., Nat. Cent. Univ., Taoyuan, Taiwan, 2022.

[21] A. T. Damaliana, A. Muhaimin, and D. A. Prasetya, "Forecasting the Occupancy Rate of Star Hotels in Bali Using the XGBoost and SVR Methods," *Journal of Statistics*, vol. 12, Jun. 2024.

[22] N. Barizki, "Daftar Harga Rumah Jabodetabek," Kaggle. Accessed: Oct. 11, 2024. [Online]. Available: https://www.kaggle.com/datasets/nafisbarizki/daftar-harga-rumah-jabodetabek/data

[DOI: 10.14710/JSUNIMUS. 12.2.2024.1-11]