Alinteri J. of Agr. Sci. (2021) 36(1): 583-589
e-ISSN: 2587-2249
info@alinteridergisi.com

http://dergipark.gov.tr/alinterizbd
http://www.alinteridergisi.com/
DOI:10.47059/alinteri/V36I1/AJAS21081

RESEARCH ARTICLE

Tree based Machine Learning in Predicting the Price of Green Building

Suraya Masrom[1*] • Thuraiya Mohd[2] • Nur Syafiqah Jamil[3]

[1*]*Associate Professor, Machine Learning and Interactive Visualization (MaLIV) Research Group, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch Tapah Campus, Perak Malaysia, Malaysia.*

[2]*Associate Professor, GreenSafe Cities (GreSAFE) Research Group, Faculty of Architecture, Planning, & Surveying Universiti Teknologi MARA, Perak Branch, Perak Malaysia, Malaysia.*

[3]*Graduate Research Assistant, Faculty of Architecture, Planning, & Surveying, Universiti Teknologi MARA, Perak Branch, Perak Malaysia, Malaysia.*

ARTICLE INFO

Article History:
Received: 04.04.2021
Accepted: 05.05.2021
Available Online: 21.06.2021

Keywords:
Tree based Machine Learning
Green Building
Price Prediction

ABSTRACT

Researchers and industry players acknowledged that machine learning application is useful in assisting human for solving many kinds of real life problems, including in real estate and property industry. In this paper, we present the empirical steps for implementing machine learning approaches in the prediction of green building price. Green building conserve natural resources and reduce the negative impact of the building development. This paper provides a report from the data collection method, preliminary data analysis with statistical method, and the experimental implementation of the machine learning models from training, validating to testing. The results show that the tree based machine learning produced better performances on the green building properties, which further tested with another five hold-out data. The testing results show that the machine learning with tree based scheme was able to predict the green building price higher than the observed price for the eight out of the ten cases within the acceptable valuation ranges.

Please cite this paper as follows:

Masrom, S., Mohd, T. and Jamil, N.S. (2021). Tree based Machine Learning in Predicting the Price of Green Building. *Alinteri Journal of Agriculture Sciences, 36*(1): 583-589. doi: 10.47059/alinteri/V36I1/AJAS21081

Introduction

A green building (GB) is beneficial to reduce the properties development negative impact on human health and environment by improving the complete building life cycle in each operation, maintenance, design, construction and evacuation [1]. Accommodations and workplace with GB specifications are mostly depending on the consumption of as much as natural resources such as increasing the utilization of water, energy, and green materials [2]. The functionality of green building does not only improve the surrounding environment, but it also enables those living and working inside the buildings to appreciate healthier atmosphere, which is free from unnecessary waste and pollution [1][3].

* Corresponding author: Suraya Masrom

The Malaysian government with the related properties agencies are actively promotes green technologies towards implementing sustainability practices in Malaysia through the National Green Technology Policy. The emerging of GB market impacts the overall property market, specifically on capital and rental values[4]. To date, a variety of methods can be used to evaluate the building valuation and the most interesting way is by using machine learning, a subfield of artificial intelligence technology.

Despite the rapid progress of machine learning utilization, none of literature can be found on the investigation of GB valuation with the machine learning approach. What is the relationship of factors that affect the GB price and how to implement the machine learning model for the GB are the two research questions that need to be answered in this research. As the research used raw collected data, it is important to conduct statistical analysis to understand the relationship of GB factors, which will be used in the machine learning prediction model. All the

empirical steps are described in the methodology part of this paper.

This paper presents useful fundamental knowledge on the design and implementation of machine learning models with tree based machine learning namely as Random Forest Regressor and Decision Tree Regressor in predicting the price of GB.

Research Background

Green Building (GB)

GB is defined as a building specification that enriches the construction sites and building efficiency based on the utilization of energy, water and natural materials[1][3]. In Malaysia, the Green Building Index (GBI) defined by Malaysia Green Building Confederation (MGB) has been used to identify the GB categorization, which rating as Platinum, Gold, Silver, and Certified. Every GBI rating has its own points, which the highest point is on the Platinum rating.

The GBI rating influences the building prices and used as a significant guide for the developers and owners in assisting the GB current prices[4]. Other criteria than rating for GB valuation are water and energy efficiency, indoor environmental quality, good connection with public transport and security[5].Physical structure is also important to consider in order to ensure the quality and efficiency of the property's performance, neither GB or non GB[6]. The fundamental aspects of physical structure includes number of bedrooms, building age, main floor area, lot size, roofing materials, floor types, ceiling types and building material [7]–[9].

Additionally, location is a major determinant of a non GB property prices, but expected to be relevant to GB also. Furthermore, the access to amenities [10], accessibility distance to public transport[11], proximity to Central Business District(CBD), closes neighborhood, schools and kinder garden, traffic connection, shop and health institution are another significant factors identified in[12].

Real Estate Valuation

Existing works on real estate valuation employed different types of conventional valuation approaches, including residual method, profit method, cost method and Discounted Cash Flow (DCF). Each approaches has their own strengths and limitations and relies on the property categories as well and the research is keep on progressing on exploring the latest and more efficient evaluation techniques. Furthermore, machine learning prediction models are seen to be beneficial, hence have been started to be used for the real estate valuation. However, there is no research has reported the use of machine learning models in the analyzing GB prices. The common machine learning modelling techniques that are already being implemented in real estate problems are Linear Regression, Decision Tree, Random Forest, Ridge Regression and Lasso Regression[13][14][15][16].Decision Tree and Random Forest are categorized as Tree Based Machine Learning[17].

Methodology

Preliminary Data Analysis

The dataset is a collection of GB housing prices in 2018 from the Valuation and Property Management Department (JPPH) Malaysia, for condominium property in the district of Kuala Lumpur. Table 1 summarizes the selected variables for the prediction model.

Table 1. The Variables of the Model

Variable	Characteristic	Machine learning feature code	Description
Dependent Variable (DV)		Transaction Price	Transaction Price (RM)
Independent Variable (IV)	Structural	LA	Lot Area
		MFA	Main Floor Area
		Tenure	Tenure
		TOP	Type of Property
		AOB	Age of Building
		NB	Number of Bedrooms
		FLOOR	Building Floor
		LPU	Level Property Unit
		Facade	Building Facade
	Location	Distance	Distance to CBD
		Access	Accessibility
		Mukim	Mukim
	Environment (Green Building)	Green Building	Green Label Certification
	Transacted Related	DOT	Date of Transaction
	Neighborhood	PD	Population density
		SEC	Security
		ID	Infrastructure development

The selection of 18 variables for the model represents several characteristics grouped by structural characteristics, location, environment, transacted and neighborhood. The data were derived from the published literature provided by JPPH as well as through the site inspection process. The original collected data contained a record of 1858 of GB and non-GB transactions under the condominium categories in the district of Kuala Lumpur, Malaysia. However, after the data cleaning process, only 240 GB transactions were involved. Table 2 shows the data cleaning process.

Table 2. Data Cleaning

No	Data	Number of Record Removed	Number of Record Left
1	Available data 2018 for Kuala Lumpur district	0	1858
2	Excluding another residential category property	122	1736
3	Remove Incomplete Data	0	1736
4	Remove Transaction from Developer	70	1666
5	Remove Non-Green Building	1426	240
6	Green Building that registered with Green Building Index in Kuala Lumpur District	0	240

From the 240 filtered data, the distribution of GBI is presented in the Table 3, which mostly were indexed as certified (81.7%).

Table 3. The Distribution of GBI rating

GB	Frequency	Percent
CERTIFIED	196	81.7
SILVER	13	5.4
GOLD	30	12.5
PLATINUM	1	0.4
Total	240	100.0

It is important to examine the distribution of data for all variables from the samples of data. Based on the normality Shapiro-Wilk test on the sample size (n=240), the researcher can identify that the distributions of all variables, which are not normally distributed (P-value < 0.05) as depicted in Table 4.

Table 4. Normality Test of All Features

	Shapiro-Wilk			
	Statistic	df	Sig.	Decision
Transaction Price	.528	240	.000	Distribution not normally distributed (p-value = 0.000 < 0.05)
Date of Transaction (DOT)	.959	240	.000	Distribution not normally distributed (p-value = 0.000 < 0.05)
Main Floor Area (MFA)	.780	240	.000	Distribution not normally distributed (p-value = 0.000 < 0.05)
Lot area	.780	240	.000	Distribution not normally distributed (p-value = 0.000 < 0.05)
No of Bedroom (NB)	.810	240	.000	Distribution not normally distributed (p-value = 0.000 < 0.05)
Level Property Unit (LPU)	.973	240	.000	Distribution not normally distributed (p-value = 0.000 < 0.05)
Building Floor	.825	240	.000	Distribution not normally distributed (p-value = 0.000 < 0.05)
Age of Building (AOB)	.798	240	.000	Distribution not normally distributed (p-value = 0.000 < 0.05)
Distance	.918	240	.000	Distribution not normally distributed (p-value = 0.000 < 0.05)
Population Density (PD)	.800	240	.000	Distribution not normally distributed (p-value = 0.000 < 0.05)

Since the distribution of data is not normally distributed, Spearman Rank correlation coefficient was implemented to identify the relationship between the transaction price (DV) and the features (IVs) as presented in Table 5.

Table 5. Summary of Result from the Spearman's Rank Correlation Analysis

	DV	DOT	MFA	Lot Area	NB	LPU	FLOOR	AOB	Distance	PD
DV	1.0	-.103	*.789***	*.788***	*.541***	.094	*.423***	.079	*-.374***	*.692***
DOT		1.0	-.092	-.091	-.191***	-.021	-.084	-.015	-.031	-.023
MFA			1.0	1.00***	.513***	-.166*	.099	.413***	-.240***	.452***
Lot Area				1.0	.513***	-.167***	.097	.414***	-.241***	.451***
NB					1.0	.103	.446***	.063	-.188***	.278***
LPU						1.0	.333***	-.091	.063	.005
FLOOR							1.0	-.004	-.402***	.543***
AOB								1.0	.033	-.125
Distance									1.0	-.731***
PD										1.0

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

DV-Transaction Price

The result indicates that the selected variables used in this research have some degree of relationship. There is a strong positive linear correlation (rho from 0.692 to 0.789 at p-value <0.01) between transaction price with lot area, main floor area (MFA), and population density (PD). Moderate correlation can be found with number of bedrooms (NB) (rho=0.541, p-value=0.000 < 0.01). On the contrary, there is a weak positive linear correlation between transaction price and building floor (FLOOR) (rho=0.423, p-value=0.000 < 0.01). A weak negative correlation can be seen between the DV and distance (rho = -0.374, p-value < 0.01). However, there is not enough evidence to support that there is a relationship between date of transactions (DOT), level of property unit (LPU) and age of building (AOB) with the DV.

As explained by the results in Table 5 , there exists variables that do not show a bivariate correlation to the DV, but they may show relation in the regression with the influences of other variables. Based on ANOVA testing, we found that all the IVs including the no show bivariate correlation variables, have contributed some degree of information at R squared 83.40% of the total variation in the transaction price. Table 6 presents the ANOVA result.

Table 6. ANOVA Result

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	549715548267383.75	10	54971554826738.375	114.74	0.000
Residual	109708947536789.64	229	479078373523.099		
Total	659424495804173.40	239			

*R^2 = 0.834

The next step is to answer the research question two on the implementation process for the machine learning model. We use Jupyter Notebook to implement the Python program for all the models. The hardware specification is Intel i7 7th Generation processor and 16 GB RAM. Based on the preliminary data analysis, we decided to break up the experiments into two groups of features selection. First group involved all the independent variables and the second group only used the bivariate correlation variables as depicted in Figure1.

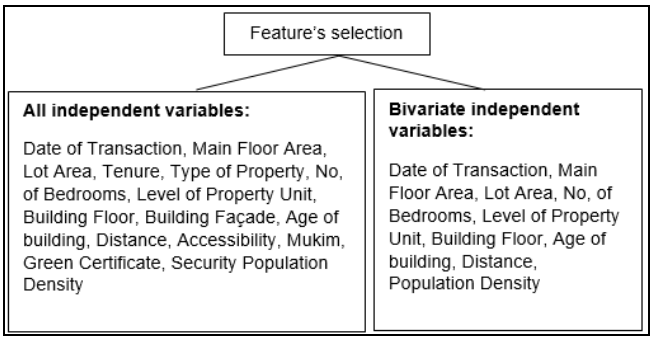

Figure 1. Features selection for the machine learning model in two different experiments

Training, Validating and Testing the Machine Learning

Training is an important task to give relevant information, knowledge, or data pattern exposure about the prediction dataset to be learned by the machine learning algorithms. Based on the given training dataset, each algorithm will study the pattern of data based on predefined calculation designed to them. Based on the pattern, each machine learning algorithm can later use the information to make decision for prediction. In the end of the process, if the machine learning is given with a new holdout sample, they will be able to infer or make decision to produce the prediction result according to the gained experienced that it learns from the pattern of training dataset. To conduct the training, validating and testing on the collected dataset, three division of dataset was given to the machine learning as presented in Figure 2.

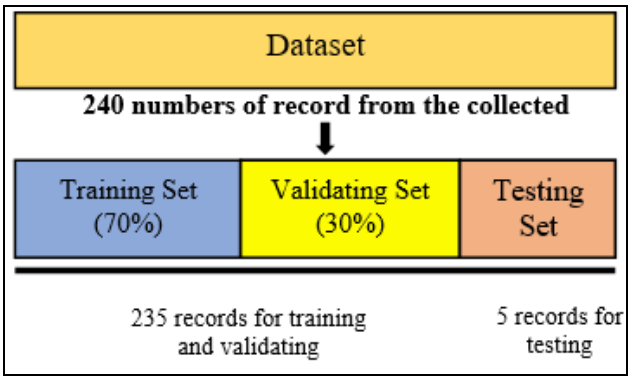

Figure 2. Training, Validating and Testing Dataset

From the 240 numbers of records, 5 records were used for testing, which manually removed from the collected datasets and saved in another excel file. Furthermore, the rest of 235 records (saved in one Excel file) were used for training and validating with division of 70:30 ratio of percentages. From the 70% of 235 data-sets, 164 rows of records were used for training and the rest 71 were used for validating. The technique used in dividing the collected dataset for training, validating, and testing purpose was categorized as Split approach.

The Machine Learning Algorithms

This study used five machine learning algorithms that are commonly used as building price predictors, divided into two paradigms, namely tree and linear. Two algorithms for tree paradigm are Random Forest Regressor and Decision Tree Regressor. Ridge, Lasso and Linear Regressor are the linear based machine learning.

At the early stage, auto hyper-parameter tuning was implemented based on the training dataset by calling *best_estimator* from the Python Scikit-Learn library. The best estimator employs grid search optimization of hyper-parameter tuning on the given machine learning algorithm. Figure 3presents the steps to implement the best estimator in the training and validating processes.

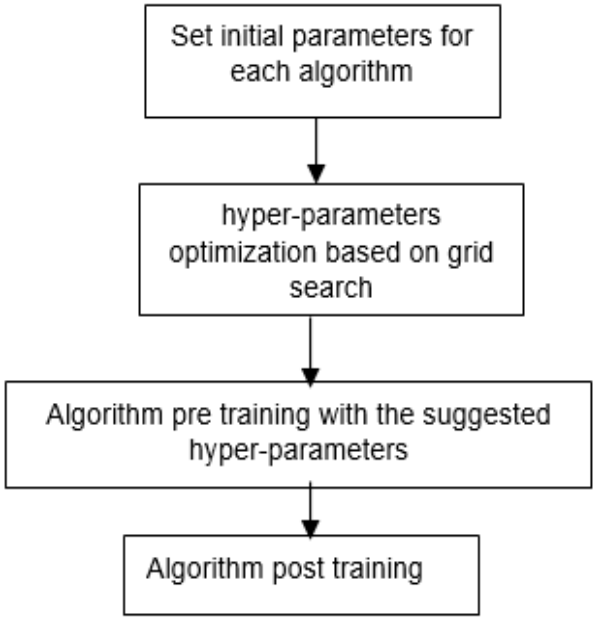

Figure 3. Hyper-parameter tuning

Based on the best estimator parameters tuning, the optimal parameters for random forest is 20 number of trees with maximal depth 7. For the decision tree, the optimal parameter suggested is on the maximal depth, which was set as 10. After the post training, GB price prediction were implemented where the accuracy and error rate were recorded for comparison.

Testing the Machine Learning Models

During the testing phase, the predicted transaction prices produced by the tree based machine learning will be compared with the observed transaction price. The reason for testing is to observe how accurate the tree based model in predicting the target variable when it is given with new datasets (hold-out sampled) that are not exposed to them during the training process. Data for testing was randomly selected from the collected data, which has been initially separated.

Results

The results in Table 7 presents the average R^ and RMSE from the average of five times experiments of each machine learning algorithms based on different selection group on the GB dataset.

Generally, the Decision Tree regressor produced higher R^ values and lower RMSE compared to other algorithms followed by the Random Forest regressor in both group of features selection. Lasso and Linear Regression worked at moderate performances with R^ between 0.81 to 0.84. Ridge model has the lowest R^ and the highest RMSE from all of the models. Therefore, the tree based machine learning is outperformed the linear algorithms.

Table 7. Performances of each algorithm based on two features selection group

	Features selection 1		Features selection 2	
	R^	RMSE (%)	R^	RMSE (%)
Decision Tree	0.944	13.42	0.921	13.48
Random Forest	0.914	16.64	0.898	16.67
Ridge	0.766	27.43	0.756	27.44
Lasso	0.841	22.59	0.841	22.59
Linear Regression	0.812	24.57	0.811	24.59

Focus to the features selection strategy, it can be seen that all independents variables are needed to the prediction models for all algorithms. It can be seen a slight performances decrement when the machine learning models only used the bivariate correlation in features selection 2.

The tree based machine learning with Decision Tree and Random Forest algorithm were selected to be tested as they produced the highest R^ and lowest RMSE compared the other models. By using 5 hold-out data randomly selected from the sample, the predicted prices generated by the two models were compared with the observed prices as listed in Table 9.

Table 9. The testing results

No	Observed Price	Predicted Price	RMSE (%)
Decision Tree			
1	688000	730000	21.63
2	1750000	1230000	23.30
3	1080000	1150000	25.81
4	473000	610000	18.48
5	834300	730000	16.61
Random Forest			
1	688000	773222	23.63
2	1750000	1464533	14.86
3	1080000	1347187	29.21
4	473000	680500	11.11
5	834300	850386	6.22

Observed price is the actual market price, whereas predicted price is the predicted result of the tree based machine learning models. The predicted results from the Decision Tree and Random Forest were mostly higher than observed prices. As presented in Figure 4 and Figure 5, most of the prices predicted by the both models are higher than the observed prices. To value a building price higher than the market price is a good strategy in the real practice of house marketing.

Figure 4. Decision Tree Model for Green Building Condominium

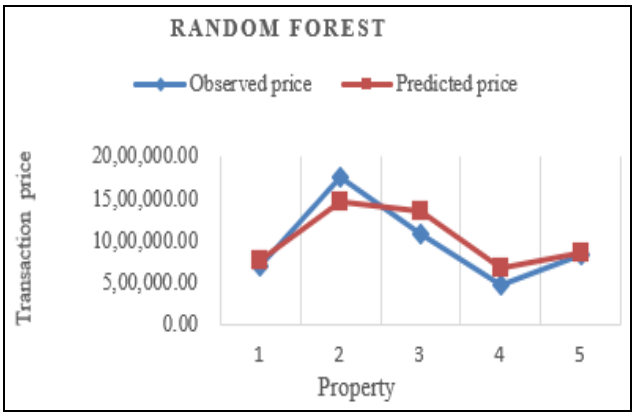

Figure 6. Random Forest Model for Green Building Condominium

Therefore, the two machine learning models with Decision Tree and Random Forest algorithms can be considered as accepted models for predicting a GB building price.

Results

Within the scope of this study, it shows that the features used in this study are signification to the building price prediction. However, as this is the first attempt of machine learning model for the GB price prediction model, the study needs a lot of improvement in future works. There are still many other features that can contribute to the property price should be involved in the model. Among the five selected machine learning algorithms, researcher decided to choose Decision Tree and Random Forest model to be tested with the new datasets. In future works, researchers should employ other kinds of machine learning algorithms for the prediction and try to test the models with the dataset on GB from other areas in Malaysia. This paper presents a set of empirical experiments demonstrating a hyper-parameter approximation approach using Scikit-Learn Python library. Future works may consider automated machine learning (AutoML) that used other kind of intelligent and deterministic algorithms such as Particle Swarm Optimization and Genetic Programming.

Acknowledgments

The authors would like to thank UNIVERSITI TEKNOLOGI MARA and Ministry of Education Malaysia for the financial support to this project under FRGS grant No 600-IRMI/FRGS 5/3 (208/2019).

References

Radwan, M.R., Kashyout, A.E.H.B., ELshimy, H.G., and Ashour, S.F., 2015. Green building as concept of sustainability Sustainable strategy to design Office building. *2 nd ISCASE-2015 Dubai*, *41*.

Darko, A., Chan, A.P.C., Owusu, E.K., and Antwi-afari, M.F. 2018. Benefit of Green Building: A Literature Review. *Convererence RICS COBRA* 5

Ismail, N., Rahmat, M.N., and Said, S.Y., 2015. Proceedings of the Colloquium on Administrative Science and Technology. In *Proceedings of the Colloquium on Administrative Science and Technology* 311-323. https://doi.org/10.1007/978-981-4585-45-3

Rashid, A., 2019. *The Equation Model of Transaction price for Green Building Certificate on the Condominium In Selangor*. Univ. Teknol. Mara campus Shah Alam Selangor Darul Ehsan, 1–93.

Dziauddin, F., 2014. The determinants of house prices in the Klang Valley, Malaysia. *Perspektif: Jurnal Sains Sosial dan Kemanusiaan*, 6(1): 70-80.

Yusof, A. M., and Ismail, S., 2012. Multiple regressions in analysing house price variations. *Communications of the IBIMA*, 1-9.

Jayantha, W.M., and Man, W.S., 2013. Effect of green labelling on residential property price: a case study in Hong Kong. *Journal of Facilities Management*, 11(1): 31–51.

Mohamad, J., 2012. *Transformation Regression and Multiple Regression Analysis*. Univ. Teknol. Malaysia, 1–107.

Rahadi, R.A., Wiryono, S.K., Koesrindartoto, D.P., and Syamwil, I.B., 2015. Factors influencing the price of housing in Indonesia. *International Journal of Housing Markets and Analysis*, 8(2): 169-188.

Kamal, E. M., Hassan, H., and Osmadi, A., 2016. Factors influencing the housing price: developers' perspective. *International Journal of Humanities and Social Sciences*, 10(5): 1676-1682.

Ferlan, N., Bastic, M., and Psunder, I., 2017. Influential factors on the market value of residential properties. *Engineering Economics*, 28(2): 135-144.

Candas, E., Kalkan, S.B., and Yomralioglu, T., 2015. Determining the factors affecting housing prices. *In FIG Working Week*, 17-21.

Shinde, N., and Gawande, K., 2018. Valuation of house prices using predictive techniques. *International Journal of Advances in Electronics and Computer Science*.

Mohd, T., Masrom, S., and Johari, N., 2019. Machine learning housing price prediction in petaling jaya, Selangor, Malaysia. *International Journal of Recent Technology and Engineering*, 8(2 Spec): 542-546.

Ma, Y., Zhang, Z., Ihler, A., and Pan, B. (2018). Estimating warehouse rental price using machine learning techniques. *International Journal of Computers Communications & Control,* 13(2): 235-250.

Ravikumar, A.S., 2017. *Real Estate Price Prediction Using Machine Learning* (Doctoral dissertation, Dublin, National College of Ireland).

Rahmati, O., Falah, F., Naghibi, S.A., Biggs, T., Soltani, M., Deo, R.C., and Bui, D.T., 2019. Land subsidence modelling using tree-based machine learning algorithms. *Science of the Total Environment*, 672: 239-252.