Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur

Shuzlina Abdul-Rahman[1], Sofianita Mutalib[4]
Research Initiative Group of Intelligent Systems
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia

Nor Hamizah Zulkifley[2]
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia

Ismail Ibrahim[3]
Data Science Department
PETRONAS Digital Sdn Bhd
Kuala Lumpur, Malaysia

***Abstract*—House price is affected significantly by several factors and determining a reasonable house price involves a calculative process. This paper proposes advanced machine learning (ML) approaches for house price prediction. Two recent advanced ML algorithms, namely LightGBM and XGBoost were compared with two traditional approaches: multiple regression analysis and ridge regression. This study utilizes a secondary dataset called 'Property Listing in Kuala Lumpur', gathered from Kaggle and Google Map, containing 21984 observations with 11 variables, including a target variable. The performance of the ML models was evaluated using mean absolute error (MAE), root mean square error (RMSE), and adjusted r-squared value. The findings revealed that the house price prediction model based on XGBoost showed the highest performance by generating the lowest MAE and RMSE, and the closest adjusted r-squared value to one, consistently outperformed other ML models. A new dataset which consists of 1300 samples was deployed at the model deployment stage. It was found that the percentage of the variance between the actual and predicted price was relatively small, which indicated that this model is reliable and acceptable. This study can greatly assist in predicting future house prices and the establishment of real estate policies.**

Keywords*—House price; house price prediction; machine learning; property; regression analysis

I. Introduction

House is one of the most essential basic needs in human life, along with other basic needs such as food and water. Demand for houses has rapidly increased through the years as people's standard of living has improved. Even though there are people who make their house as an investment and asset, yet most people around the world buy a house to live in. Undoubtedly, the housing sector has a positive impact on a country's currency, which is an important scale for the national economy [1]. Homeowners will buy goods such as furniture and house appliances for their house and home builders or contractors will buy raw material to make houses to fulfil the demand for houses, which is an example of the economic wave effect created from new house supply. Meanwhile, consumers have the capital to make a large investment, and the construction industry is vibrant or otherwise can be seen through the high level of house supply or demand in a country. Nevertheless, house has become unaffordable as there is a significant price expansion in the housing market sector in many countries [2].

In Malaysia, buying a house is never an easy experience because this decision can cost a lot of money. According to the Department of Statistics Malaysia, the average price of a house in the 2nd quarter of 2018 was RM 408,774 compared to ten years ago which was RM 199,431 [3]. The average house price in Malaysia was obviously two times higher in 2018 compared to 2009. A wide variety of factors may affect house prices such as the facilities provided, type of houses, number of bedrooms and size of a house. These factors also vary depending on the location of the house, for example, there will be an obvious difference in house prices, in Singapore compared to Malaysia. Yet, we cannot simply say that the price of each house is similar throughout Malaysia as the prices of houses in Kuala Lumpur (urban area) is not the same as the price of houses in Perlis (rural area). Yop [3] in his report mentioned that the average house price in Kuala Lumpur in 2018 was RM 772,398, while the average house price in Perlis for the same year was just RM 177,945. It is suggested that the prediction of house prices could be more accurate if the prediction is considered based primarily on a specific region.

House prices can be predicted using machine (ML) algorithms including support vector regression, artificial neural network, and linear regression. The ML model that provides the best prediction results will be beneficial for researchers, home buyers, property investors, and house builders in terms of gaining a lot of knowledge and information of the house price values in the present sector. Additionally, this model can facilitate potential buyers to determine the characteristics of a house they prefer that adhere to their budget [4]. Prediction of house prices in Kuala Lumpur would be a significant research as Kuala Lumpur is the capital city of Malaysia that offers a range of facilities including efficient public transportation, shopping malls, and many more compared to states in a rural area such as Perlis has yet to provide. All the facilities provided can encourage many home buyers, investors, and house builders to supply more houses in this area, at the same time the demand for houses in this specific region would increase. Nonetheless, the suitable model in predicting house prices in Kuala Lumpur remains unclear as there is limited study conducted in this region.

Research showed that ML algorithms have proven extremely useful for addressing many predictions and classification problems with broad application scope, including customer classification and segmentation [5], market analysis [6], [7], and education [8]. Unfortunately, ML is relatively limited and very far from being used in real estate applications mainly in the Malaysian sense. Evaluation of property prices and values is extremely critical for the real estate sector, the stock market, the economics and tax sector, as well as the scale of buyers' and sellers' wallets [9]. Although researchers are relatively aware of the existing current prediction model, consideration should be given to the current methodologies constrained by the scope of data of the current system in the real estate industry. Therefore, it is important to examine the correlation between house prices and housing attributes and identify significant variables that are essential to the use of the ML techniques in the real estate industry, involving pre-processing and exploration of the datasets obtained.

As various factors such as location and property demand could affect house prices, most parties involved including buyers and investors, housebuilders, and real estate market may want to know the exact attributes or the main factors affecting house prices to assist investors in making decision and to facilitate house builders in setting house prices [2], [10], [11]. Nevertheless, other characteristics including distance from local facilities and the physicality of a house might be overlooked. This may contribute to the creation of a house price model that does not reflect the actual condition of the housing sector. In comparison, houses with identical attributes may be priced at varying rates, while houses with different type of attributes can be priced at the same amount. Besides that, real-estate industry in Malaysia is far behind compared to that in the United States and United Kingdom in valuation [2], [12] as Malaysia is still using the traditional method to valuate a house. Developers or real-estate agencies will send valuers to each house, to appraise the house price resulting in wastage of time and cost, also the fluctuation of house prices. However, this issue and many other pressing industrial problems can be effectively addressed using big-data technologies that includes ML in this age of Industry 4.0.

This paper presents an exploration of ML algorithms for house price prediction by focusing on Kuala Lumpur housing data. The study aims to propose an advanced ML algorithm that can generate a promising model for house price prediction. To achieve this aim, four models namely multiple regression analysis, ridge regression, light gradient boosting machine (LightGBM), and XGBoost are used to learn about the relationship between the house attributes and house prices as well as to predict the house price. The remainder of this paper is organized as follows: Section II describes related works on prediction of house price in Malaysia and followed by housing price prediction models. Subsequently, in Section III and Section IV, the experimental designs and the data modeling are presented consecutively. Section V presents the model deployment while Section VI discusses the results and findings of the study. Finally in Section VII, the study is concluded.

II. Related Work

A. Prediction of House Price in Malaysia

Prior to 2019, there were several studies regarding the prediction of house prices in Malaysia. In 2018, a study has been conducted by Yap and Ng [13] to determine house affordability in Malaysia. This study found out that key factors affecting affordability of houses were income, price of a property, land cost, policy of supply and demand, and changes in the economy. Their study [13] offered detailed insights into exploring housing market in Malaysia; however, they only provided a descriptive analysis, which lacked in the predictive analysis towards the housing market in Malaysia. The study by [2] found out that there were several attributes that play a significant role in determining house prices in Petaling District. Another significant finding from this study was Puchong and Petaling Jaya was classified as less volatile housing markets compared to Sungai Buloh. A similar study conducted by Abdullahi et al. [14] using Multiple Regression Analysis and Hedonic Regression Analysis in explaining price variations in Malaysia found that one of the most influential attributes was the location of a house.

Apart from that, building area and building age were significant in variations of prices of houses in Malaysia. The locations mentioned in the study were based on states in Malaysia; hence location is the most dominant attribute compared to others (attributes). Meanwhile, this study focused on the location of a small region of Kuala Lumpur to predict the house price. Thus, location cannot be generalized as the main significant attribute in predicting house prices. Another similar study conducted examined the prediction of house prices in Petaling District Malaysia. The study by Chang et al. [9] used a different method called the Functional Relationship Model. The model was used to identify the impact of residential property attributes on house price in Petaling District. Several attributes were identified such as building size and bedroom numbers which had a significant effect in predicting the house prices. The study successfully developed a new predictive model and then applied on the Petaling District terrace-houses only, even though, there existed various types of houses in the area such as bungalows, condominiums, or apartments. Taking everything into account, the model was unable to reflect the whole housing market in Petaling District.

B. Housing Price Prediction Models

ML algorithms to develop house price prediction models have been actively researched and models are constructed by using algorithms such as random forest, decision tree, lasso, and linear regression [12], [15]. A study by Wu et al. [12] categorized models in analyzing the real-estate market into the conventional valuation system and the advanced valuation system. The traditional valuation system includes the multiple regression method and stepwise regression method, while the advanced valuation is the hedonic pricing method, artificial neural network (ANN), and spatial analysis method. The choice of a model that needs to be used to predict house prices is quite critical as there is a variety of models available. One of the most widely used models in the real-estate field is the regression analysis namely the multiple linear regression, support vector regression, and hedonic regression analysis, which is used by several researchers including [9], [16]–[20]. In addition, several other machine learning models such as the gradient boosting model including Catboost, XGBoost and LightGBM, random forest, decision and artificial neural network have been used frequently in the study of real- estate [10], [11], [21]–[23]. In this study, four models namely multiple regression analysis, ridge regression, LightGBM, and XGBoost are used to learn about the relationship between house attributes and house prices as well as to predict house prices. The next subsections explain further these four models.

- Multiple Regression Analysis

Multiple Regression analysis is an extension of simple linear regression to predict the value of a variable based on the value of two or more other variables. It can determine the strength of the relationship between an outcome (the dependent variable) and several predictor variables as well as the importance of each of the predictors to the relationship [24] [25]. Four basic assumptions need to be fulfilled to use the multiple regression analysis model [26] as cited in [27]. The first assumption is the variable used in the model must be normally distributed. Multivariate data cleaning is also an important consideration in multiple regression.

The second assumption that needs to be met is the relationship between the dependent and independent variables which must be linear to estimate the variables accurately. Next, the assumption to use this model is no perfect multicollinearity between variables. The last assumption to be fulfilled is little or no auto correlation. There are various other assumptions for this model; however, these assumptions are among the easiest to deal with if needed. The prediction of house prices using multiple regression model is conducted by assigning the price of a house as a target variable or dependent variable, while other attributes are set as independent variables to determine the most significant variables by identifying the correlation coefficient of each attribute.

- Ridge Regression

Ridge regression is a technique used to assess the presence of collinearity in multiple regression data [28]. As multicollinearity happens, the estimates of the least-squares are unbiased, but their variances are large enough such that they can be far from the real value. Ridge regression provides a more credible performance by reducing the standard errors. There were several industries that have deployed ridge regression model as their solution. For example, in medical industry, this model was deployed in healthcare analysis system and blood-base tissue gene expression as well as in wind speed forecasting [29]–[31]. According to Manasa et al. [10], the ridge regression model is a regularization model that incorporates and optimizes an additional variable (tuning parameter) to resolve the effect of multiple variables in linear regression, typically referred to as ‘noise’ in a statistical sense. This model has been used by a lot of researchers in the real-estate previous studies including [10], [32]–[34]. In their study, [33] found that ridge regression is able to provide the lowest MSE value for the house prediction model compared to other models in that study including lasso and gradient boosting.

- Light Gradient Boosting Machine (LightGBM)

Initially launched in late 2017, the Light Gradient Boosting Machine or LightGBM has a stable release in November 2020. This model has been used by many researchers in their fields such as medical fields [35], and biochemistry [36], [37]. This model, however, is still rarely used in the real estate market, since there is only one academic study that has utilized this model [22]. In their research related to prediction of house prices, [22] combined this model with another prediction model, which were CatBoost and XGBoost model to forecast house rentals in China. This joint model managed to provide the smallest RMSE value of the house price prediction model.

- Extreme Gradient Boosting (XGBoost)

XGBoost is the most powerful algorithm for any regression or classification problem which stands for ‘extreme gradient booster’. The developers [38] of this model in their article describe XGBoost as a scalable tree boosting machine learning system. The framework of this model can be seen as an open-source kit. In a range of machine learning and data mining problems, the influence of this model has been widely recognized. In fact, many successful machine learning competitions utilize these styles of model. Besides, this model is often used in real-world development networks, including internet ads. According to [34], XGBoost algorithm addresses issues of the linear regression model. The XGBoost model can handle numerical as well as categorical variables very well. This model can automate different loss functions and offers many tuning choices for machine learning engineers to modify the model. This model had been used in many fields including the real estate and housing market sector [10], [32], [34], [39]. Most studies in the real estate industry that use the XGBoost model found out that this model is able to provide the lowest RMSE value for the prediction of house prices compared to other models [23], [34]. Table I summarizes past studies that work on similar domains from 2011 until 2020. As can be seen, the first four ML models appear to be the most commonly used by past researchers. The list of variables that associate with this domain can be referred in our paper [40]. As stated in the paper, the factors influencing house prices can be classified into three categories: location, structural and neighborhood condition.

TABLE I. SSUMMARY OF LITERATURES ON HOUSE PRICE PREDICTION USING ML MODELS

Previous study	**MRA**	**RR**	**LGBM**	**XGB**	**RF**	**NN**
[23]	✔			✔		
[21]			✔	✔	✔	
[10]		✔		✔		
[34], [39]	✔	✔		✔		
[16], [20]	✔					
[22]			✔	✔		
[15]	✔				✔	
[41], [42], [19]						✔
[43]	✔					
[32], [33]		✔		✔		
[14],[17], [44]	✔					

[a] MRA: Multiple Regression Analysis; RR: Ridge Regression; LGBM: Light Extreme Gradient Boosting; XGB: Extreme Gradient Boosting; RF: Random Forest; NN: Neural Networks

III. Experimental Setup

A. Data Preparation

This study is based on a secondary dataset and retrieved from the Kaggle website (https://www.kaggle.com/dragonduck/property-listings-in-kuala-lumpur) and Google Maps. The dataset has originally scrapped from property listings in Kuala Lumpur, Malaysia in 2019. The original dataset contains only 8 variables including location, price, rooms, bathrooms, property type, size, and furnishing. Other than that, there were other four variables that also showed a significant impact to house price prediction derived from Geocoder and Googleplaces Python package used to retrieve locations through Google Maps. The variables were distance to shopping mall, distance to hospital, access of public transport, and distance to nearest school. Thus, the final dataset contains 53883 observations with 12 variables including one target variable. In this study, the target variable was the price (which is a continuous variable) in Ringgit Malaysia (MYR). Meanwhile, the independent variables were location (location of a house), bedroom (number of bedrooms available), bathroom (number of bathrooms available), car park, size (house lot size), furnishing status, property type, shopping mall (the nearest shopping mall to the house in KM), school (the nearest school to the house in KM), hospital (the nearest hospital to the house in KM) and public transport (the nearest LRT and MRT station to the house in KM). Table II shows the description of these variables.

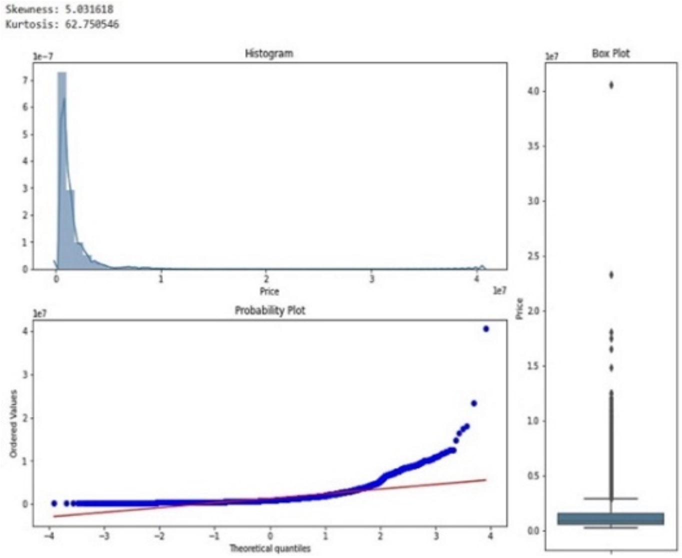
TABLE II. DESCRIPTION OF VARIABLES

No.	**Variable Name**	**Role**	**VariableType**	**Description**
1.	Price	Target	Continuous	Price of house sold in the market
2.	Location	Input	Nominal	The location of a house for example: 1. Mont Kiara 2. City Centre 3. Bangsar 4. Desa Park City 5. Bukit Tunku (KennyHills)
3.	Room	Input	Discrete	Number of rooms in a house
4.	Bathroom	Input	Discrete	Number of bathrooms in a house
5.	Car Park	Input	Discrete	Number of car parks provided for a house
6.	Size	Input	Continuous	The built-up area of a house or the land area of a house
7.	Furnishing	Input	Nominal	Furnishing status of a house: 1. Fully furnished 2. Semi-furnished 3. Unfurnished
8.	Property Type	Input	Nominal	The type of condominium: 1. Corner 2. Duplex 3. End Lot 4. Intermediate 5. Penthouse 6. SOHO 7. Studio 8. Triplex
9.	Shopping Mall	Input	Continuous	The nearest shopping mall to the house in kilometer (KM)
10.	School	Input	Continuous	The nearest school to the house in kilometer (KM)
11.	Hospital	Input	Continuous	The nearest hospital to the house in kilometer (KM)
12.	LRT/MRT	Input	Continuous	The nearest LRT or MRT station to the house in kilometer (KM)

The data preparation stage included several tasks to create the final dataset. Tasks for data preparation were likely performed several times, and not in any specified order. Tasks included selection of table, record, and attribute, data cleaning, new attribute building, and data transformation for modeling tools were performed at this stage. For this study, the original dataset of a property listing was reduced through the data cleaning process stages and data transformation. Data cleaning can be referred to as a process of identifying and correcting errors in a dataset, for example removing missing values, and data transformation refers to a process of transforming a data to be more valuable towards a study. The invalid data including missing values and inconsistent data observed in this stage. A total of 21,995 rows containing missing values in which the majority of the missing values existed in more than two columns (good if we know the variables). All 21,955 rows with missing values were removed using Python. There were 31,899 rows left after removing missing values from the dataset.

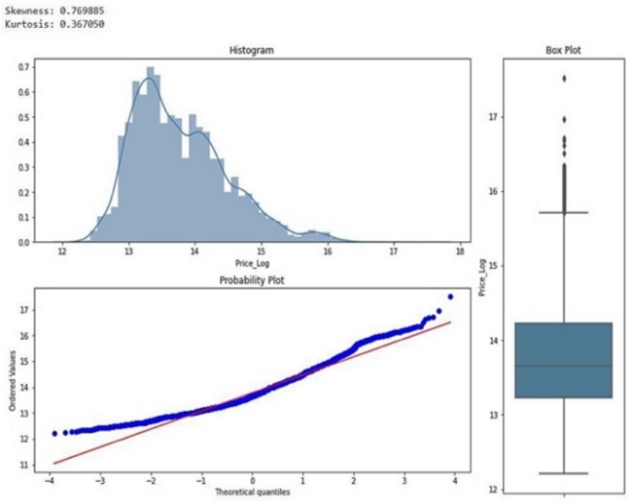
B. Data Transformation

To examine whether the dataset is normally distributed or not, the study used three graphical methods such as histograms, QQ-plot, and boxplot. On the other hand, skewness and kurtosis were also used to assess normality of each variable. According to [45], a standard normal dataset has a kurtosis value between -3 to 3. If the kurtosis value is higher or lower than this value, it will be considered as a thin bell shape. Meanwhile, [46], [47] suggested that if the skewness value is between -2 to 2, the dataset can be roughly considered as normal. Based on Fig. 1, the skewness value is 5.03 and the kurtosis value is 62.75. It means that the dataset is not normally distributed. Hence, data transformation needs to be done to make the dataset conform to normality. The transformation of the dataset was executed by using log-transformation to normalize the dataset. The transformed data showed better value of skewness and kurtosis which were 0.7699 and 0.3670, respectively. These two values are depicted in Fig. 1 and Fig. 2. The skewness value (0.77) between -2 and 2 can be considered as normally distributed.



```
Skewness: 5.031618
Kurtosis: 62.750546
```

Fig. 1. Distribution of Price before Transformation.



```
Skewness: 0.769885
Kurtosis: 0.367050
```

Fig. 2. Distribution of Price after Transformation.

C. Identification of Importance Features using Correlation Score

Pearson correlation matrix was utilized to check for the correlation value between independent and dependent variable. This test helped this study determine which attributes played an important role in valuating house prices. On the other hand, the correlation value was also used to detect whether there is multicollinearity in this dataset. According to Ratner (2009), multicollinearity values between 0 and 0.3 (0 and-0.3) imply a weak positive (negative) linear relationship via a shaky linear law, while values between 0.3 and 0.7 (0.3 and−0.7) via a fuzzy-firm linear rule suggest a moderate positive (negative) linear relationship and the values between 0.7 and 1.0 (-0.7 and -1) suggest a strong positive (negative) linear relationship. Fig. 3 illustrates the heatmap of the Pearson correlation coefficient matrix based on each variable. As depicted in Fig. 3, variable hospital which is the distance to the nearest hospital has a weak negative correlation which is -0.19. Meanwhile, other attributes show a moderate and strong correlation value. However, there are no attributes being dropped from this study, as this study utilizes machine learning model that can detect itself which attributes is good or not.

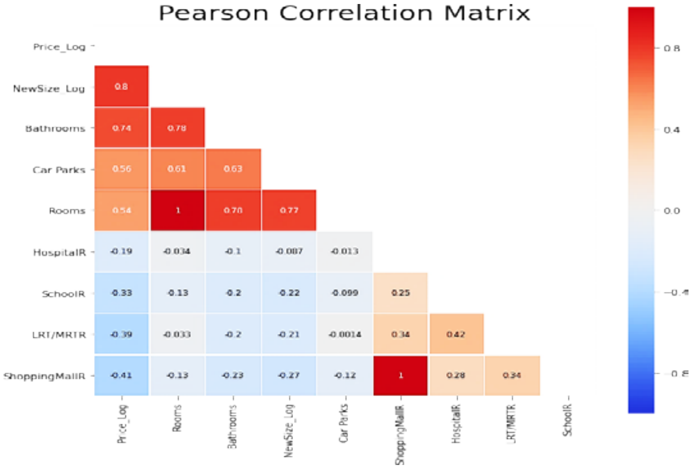


Fig. 3. Heatmap of Pearson Correlation Coefficient Matrix.

D. *Data Representation using One Hot Encoding*

One hot encoding is a technique used to express categorical features as numerical variables with more interpretable outcome values that will be easier to be understood by a machine learning model. For instance, the furnishing status is conveyed by a name that characterizes how furnished the space is (fully furnished, semi-furnished, and unfurnished). This technique transformed the input into 3 columns. Each column denotes a condition or status of furnishing. The column of a specific status of furnishing is 1, and those in the same row are set to 0.

IV. Data Modeling

In this study, predictive modeling was carried out using multiple regression analysis, ridge regression, LightGBM, and XGBoost. The dataset then was partitioned into two groups, training and testing sample. The training sample partition which consists of 70% of the dataset and 30% of the dataset was used in the testing sample. These four models were then compared in order to select the best model. This study utilized the hyper-parameter automated search module-GridSearchCV to search for the optimal parameter values to enhance the efficiency of each model [48]. Without losing generality, the RMSE value with optimal parameters can be generated in this module.

A. *Implementation of Multiple Regression Analysis*

The assumption of linear regression such as linear relationship, normal distribution of all variables, no perfect multicollinearity and little or no autocorrelation was checked before the model was used. Supposed that y represents the dependent variable, which is the prediction of house price, βk is the coefficient value of the linear function, and xk denotes each of the attributes used in this model such as the size of house and furnishing status. According to Manasa et al. [10], this model can be represented in a mathematical model as in equation (1):

$$y = \sum_{0}^{k} \beta_k x_k \qquad (1)$$

where;

x_k, k = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11 denote eleven attributes for each house, y = the predicted price of a house in Kuala Lumpur. Several assumptions regarding linear regression were checked prior to the usage of this model. The first assumption of linear regression model is linearity. This model assumes that there is a linear relationship between the independent and dependent variables. Fig. 4 shows the graph of newsize_log against price_log. As can be seen, the graph shows a non-perfect linear relationship. However, it can still be said that the independent variables have a linear relationship towards dependent variable (price of house). The second assumption of linear regression is regarding the distribution of variables. This model assumes that all variables are normally distributed. This assumption has been checked by using the Q-Q plot as illustrated in Fig. 5.

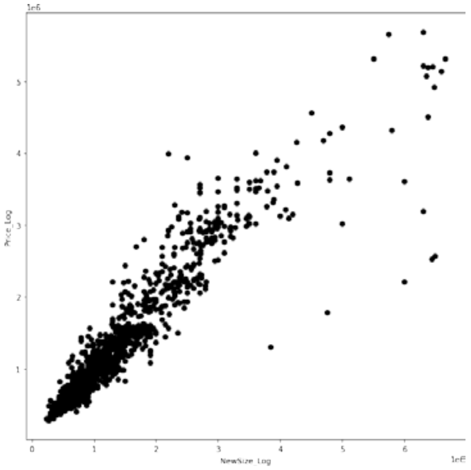


Fig. 4. Graph of Actual Price against Predicted Price.

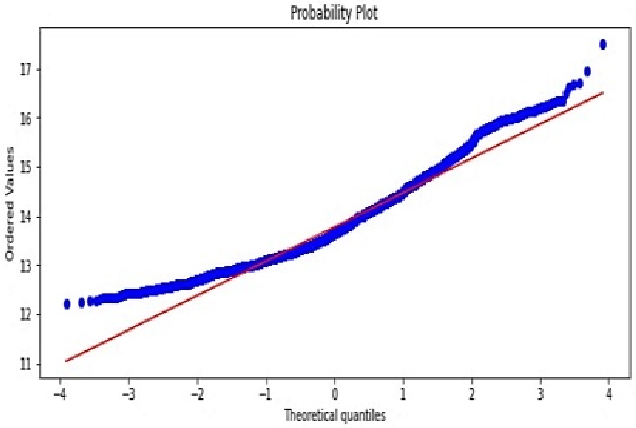
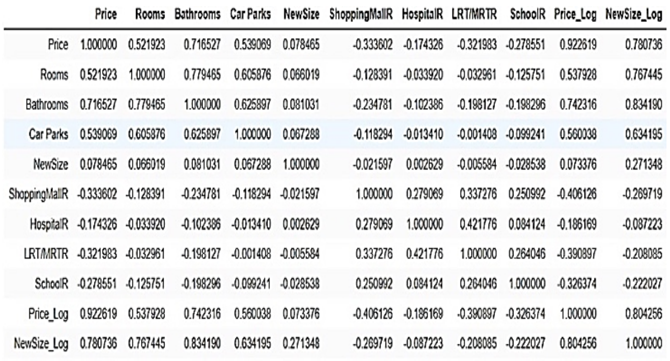


Fig. 5. The Q-Q Plot of All Variables.

Fig. 5 shows that not all the points lie perfectly on the red line which indicates that, all variables used in this study are not perfectly normal distributed. The third assumption for linear regression model is that there is no perfect multicollinearity from the variables. The multicollinearity can be examined by using the Pearson correlation score, as shown in Fig. 6.

As can be seen in Fig. 6, that there is no perfectly multicollinearity between each variable. Thus, all variables were used in this study. The last assumption for linear regression model is that there is little or no autocorrelation.



	Price	Rooms	Bathrooms	Car Parks	NewSize	ShoppingMallR	HospitalR	LRT/MRTR	SchoolR	Price_Log	NewSize_Log
Price	1.000000	0.521923	0.716527	0.539069	0.078465	-0.333602	-0.174326	-0.321983	-0.278551	0.922619	0.780736
Rooms	0.521923	1.000000	0.779465	0.605876	0.066019	-0.128391	-0.033920	-0.032961	-0.125751	0.537928	0.767445
Bathrooms	0.716527	0.779465	1.000000	0.625897	0.081031	-0.234781	-0.102386	-0.198127	-0.198296	0.742316	0.834190
Car Parks	0.539069	0.605876	0.625897	1.000000	0.067288	-0.118294	-0.013410	-0.001408	-0.099241	0.560038	0.634195
NewSize	0.078465	0.066019	0.081031	0.067288	1.000000	-0.021597	0.002629	-0.005584	-0.028538	0.073376	0.271348
ShoppingMallR	-0.333602	-0.128391	-0.234781	-0.118294	-0.021597	1.000000	0.279069	0.337276	0.250992	-0.406126	-0.269719
HospitalR	-0.174326	-0.033920	-0.102386	-0.013410	0.002629	0.279069	1.000000	0.421776	0.084124	-0.186169	-0.087223
LRT/MRTR	-0.321983	-0.032961	-0.198127	-0.001408	-0.005584	0.337276	0.421776	1.000000	0.264046	-0.390897	-0.208085
SchoolR	-0.278551	-0.125751	-0.198296	-0.099241	-0.028538	0.250992	0.084124	0.264046	1.000000	-0.326374	-0.222027
Price_Log	0.922619	0.537928	0.742316	0.560038	0.073376	-0.406126	-0.186169	-0.390897	-0.326374	1.000000	0.804256
NewSize_Log	0.780736	0.767445	0.834190	0.634195	0.271348	-0.269719	-0.087223	-0.208085	-0.222027	0.804256	1.000000

Fig. 6. The Correlation Score of Each Continuous Variable.

This last assumption was tested by using Durbin-Watson test. According to [49], there is no autocorrelation in the dataset if the Durbin-Watson test is in the range of 1.5 to 2.5. This indicates that there is no autocorrelation in this dataset. The multiple regression analysis model is being used as all the assumptions of the model have been fulfilled. The most important predictor variable analyzed by multiple regression analysis model was NewSize_Log followed by Rooms and Bathrooms which indicated the size of a house, the number of bedrooms and bathrooms. Analysis of the feature importance with Multiple Linear Regression model shows the rank of features (in sequence, highest to lowest): size of the house, number of rooms, number of bedrooms, public transport, shopping mall, carparks, school, and hospital.

B. Implementation of Ridge Regression

Equation (2) presents the equation of simple linear regression.

$$y = xb + e \quad (2)$$

where y represents the dependent variable, which is the prediction of house price, while x is the features of the matrix (number of bedrooms, location, etc.), b is the regression coefficients, and e is the residual errors. Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity that performs L2 regularization. On this basis, the variables were standardized by subtracting and dividing the respective factors by their standard deviations [10]. The rank of these five features is similar to the earlier model (i.e., Multiple Linear Regression). The first three features are NewSize_Log which is the size of a house followed by the number of bathrooms and the number of rooms. Next in the list are the nearest distance to public transport, the nearest distance to shopping mall, number of car parks, the nearest school to the house and the hospital.

C. Implementation of LightGBM

LightGBM is a Microsoft GBDT open-source algorithm. The histogram-based algorithm is used to accelerate the training process, reduces memory usage, and incorporates advanced network communication to optimize parallel learning known as the algorithm for the parallel voting tree. The leaf-wise method is used by LightGBM to find a leaf with the greatest gain in splitters. For the LightGBM algorithm, the Python LightGBM module was used to evaluate the house price dataset. The house price dataset, containing 12 attributes, was allocated randomly for 70% training and 30% testing.

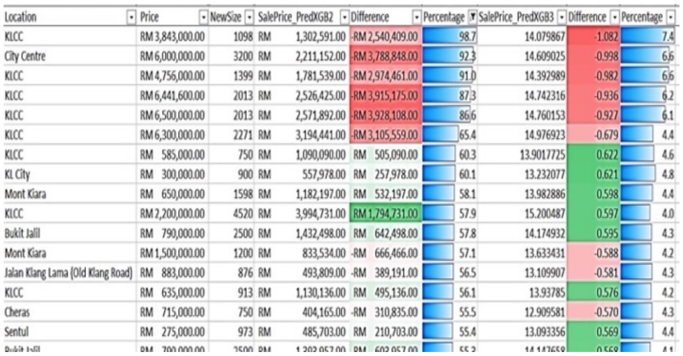
D. Implementation of XGBoost

Extreme Gradient Boosting (XGBoost) is an enhanced gradient boosting machine using the tree ensemble boosting process. This process ends in the sum of the outputs from all the trees. The XGBoost algorithm used the XGBoost package in Python to evaluate house prices. The sample data allocation scheme used in this model is the same as the previous model which is LightGBM algorithm. Analysis of the feature importance with XGBoost model shows the rank of features (in sequence, highest to lowest): size of the house, the number of parking lots provided for a house, and the nearest distance to the public transport feature. The selection of the features is quite different compared to the previous two models (Multiple Linear Regression and Ridge Regression).

V. Model Deployment

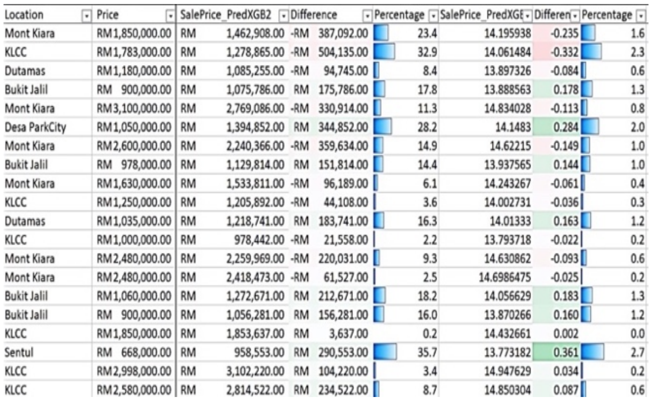
Based on the analysis results, it was recommended to deploy the XGBoost model for the prediction of Kuala Lumpur house prices using new data. This model was being executed by removing the actual house price to generate the predicted house price. To perform this task, a new dataset consists of 1300 samples of Kuala Lumpur house price was collected and our data audit results show that there are no missing values in the sample. The predicted house price was then compared with the actual house price, and the percentage difference between these two values was calculated. Fig. 7 shows the sample of the dataset which had been executed using this model. As illustrated in Fig. 7, the percentage difference for actual and predicted log price is relatively small, which indicated that this model could provide very accurate results. The percentage difference is a bit high when the log price is being inversed to get the real price.

Fig. 8 shows a few rows that contain very high values in percentage difference between predicted and actual house prices. Based on these figures, a high value in percentage difference, which is above 60%, is coming from high-end locations, for example, KLCC and KL City. It can be concluded that a house in a high-end location is difficult to be priced as it might be below or above market value. A group of highest percentage difference value which is above 80% is usually because the price of a house is under market value and is located at KLCC which is a high-end location.



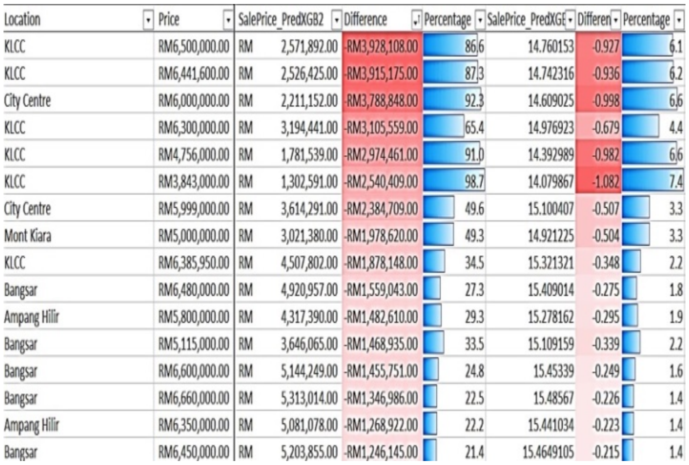
Location	Price	NewSize	SalePrice_PredXGB2	Difference	Percentage	SalePrice_PredXGB3	Difference	Percentage
KLCC	RM 3,843,000.00	1098	RM 1,302,591.00	-RM 2,540,409.00	98.7	14.079867	-1.082	7.4
City Centre	RM 6,000,000.00	3200	RM 2,211,152.00	-RM 3,788,848.00	92.3	14.609025	-0.998	6.6
KLCC	RM 4,756,000.00	1399	RM 1,781,539.00	-RM 2,974,461.00	91.0	14.392989	-0.982	6.6
KLCC	RM 6,441,600.00	2013	RM 2,526,425.00	-RM 3,915,175.00	87.3	14.742316	-0.936	6.2
KLCC	RM 6,500,000.00	2013	RM 2,571,892.00	-RM 3,928,108.00	86.6	14.760153	-0.927	6.1
KLCC	RM 6,300,000.00	2271	RM 3,194,441.00	-RM 3,105,559.00	65.4	14.976923	-0.679	4.4
KLCC	RM 585,000.00	750	RM 1,090,090.00	RM 505,090.00	60.3	13.9017725	0.622	4.6
KL City	RM 300,000.00	900	RM 557,978.00	RM 257,978.00	60.1	13.232077	0.621	4.8
Mont Kiara	RM 650,000.00	1598	RM 1,182,197.00	RM 532,197.00	58.1	13.982886	0.598	4.4
KLCC	RM 2,200,000.00	4520	RM 3,994,731.00	RM 1,794,731.00	57.9	15.200487	0.597	4.0
Bukit Jalil	RM 790,000.00	2500	RM 1,432,498.00	RM 642,498.00	57.8	14.174932	0.595	4.3
Mont Kiara	RM 1,500,000.00	1200	RM 833,534.00	-RM 666,466.00	57.1	13.633431	-0.588	4.2
Jalan Klang Lama (Old Klang Road)	RM 883,000.00	876	RM 493,809.00	-RM 389,191.00	56.5	13.109907	-0.581	4.3
KLCC	RM 635,000.00	913	RM 1,130,136.00	RM 495,136.00	56.1	13.93785	0.576	4.2
Cheras	RM 715,000.00	750	RM 404,165.00	-RM 310,835.00	55.5	12.909581	-0.570	4.3
Sentul	RM 275,000.00	973	RM 485,703.00	RM 210,703.00	55.4	13.093356	0.569	4.4
Bukit Jalil	RM 700,000.00	2500	RM 1,302,957.00	RM 602,957.00	55.2	14.147658	0.568	4.1

Fig. 7. Sample of Dataset Executed using XGBoost Model.



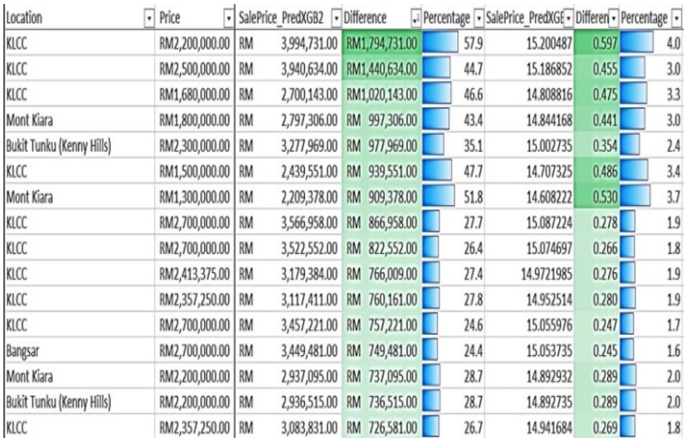
Location	Price	SalePrice_PredXGB2	Difference	Percentage	SalePrice_PredXGE	Differen	Percentage
Mont Kiara	RM1,850,000.00	RM 1,462,908.00	-RM 387,092.00	23.4	14.195938	-0.235	1.6
KLCC	RM1,783,000.00	RM 1,278,865.00	-RM 504,135.00	32.9	14.061484	-0.332	2.3
Dutamas	RM1,180,000.00	RM 1,085,255.00	-RM 94,745.00	8.4	13.897326	-0.084	0.6
Bukit Jalil	RM 900,000.00	RM 1,075,786.00	RM 175,786.00	17.8	13.888563	0.178	1.3
Mont Kiara	RM3,100,000.00	RM 2,769,086.00	-RM 330,914.00	11.3	14.834028	-0.113	0.8
Desa ParkCity	RM1,050,000.00	RM 1,394,852.00	RM 344,852.00	28.2	14.1483	0.284	2.0
Mont Kiara	RM2,600,000.00	RM 2,240,366.00	-RM 359,634.00	14.9	14.62215	-0.149	1.0
Bukit Jalil	RM 978,000.00	RM 1,129,814.00	RM 151,814.00	14.4	13.937565	0.144	1.0
Mont Kiara	RM1,630,000.00	RM 1,533,811.00	-RM 96,189.00	6.1	14.243267	-0.061	0.4
KLCC	RM1,250,000.00	RM 1,205,892.00	-RM 44,108.00	3.6	14.002731	-0.036	0.3
Dutamas	RM1,035,000.00	RM 1,218,741.00	RM 183,741.00	16.3	14.01333	0.163	1.2
KLCC	RM1,000,000.00	RM 978,442.00	-RM 21,558.00	2.2	13.793718	-0.022	0.2
Mont Kiara	RM2,480,000.00	RM 2,259,969.00	-RM 220,031.00	9.3	14.630862	-0.093	0.6
Mont Kiara	RM2,480,000.00	RM 2,418,473.00	-RM 61,527.00	2.5	14.6986475	-0.025	0.2
Bukit Jalil	RM1,060,000.00	RM 1,272,671.00	RM 212,671.00	18.2	14.056629	0.183	1.3
Bukit Jalil	RM 900,000.00	RM 1,056,281.00	RM 156,281.00	16.0	13.870266	0.160	1.2
KLCC	RM1,850,000.00	RM 1,853,637.00	RM 3,637.00	0.2	14.432661	0.002	0.0
Sentul	RM 668,000.00	RM 958,553.00	RM 290,553.00	35.7	13.773182	0.361	2.7
KLCC	RM2,998,000.00	RM 3,102,220.00	RM 104,220.00	3.4	14.947629	0.034	0.2
KLCC	RM2,580,000.00	RM 2,814,522.00	RM 234,522.00	8.7	14.850304	0.087	0.6

Fig. 8. Sample of Dataset with High Value of Percentage Difference.



Location	Price	SalePrice_PredXGB2	Difference	Percentage	SalePrice_PredXGE	Differen	Percentage
KLCC	RM6,500,000.00	RM 2,571,892.00	-RM3,928,108.00	86.6	14.760153	-0.927	6.1
KLCC	RM6,441,600.00	RM 2,526,425.00	-RM3,915,175.00	87.3	14.742316	-0.936	6.2
City Centre	RM6,000,000.00	RM 2,211,152.00	-RM3,788,848.00	92.3	14.609025	-0.998	6.6
KLCC	RM6,300,000.00	RM 3,194,441.00	-RM3,105,559.00	65.4	14.976923	-0.679	4.4
KLCC	RM4,756,000.00	RM 1,781,539.00	-RM2,974,461.00	91.0	14.392989	-0.982	6.6
KLCC	RM3,843,000.00	RM 1,302,591.00	-RM2,540,409.00	98.7	14.079867	-1.082	7.4
City Centre	RM5,999,000.00	RM 3,614,291.00	-RM2,384,709.00	49.6	15.100407	-0.507	3.3
Mont Kiara	RM5,000,000.00	RM 3,021,380.00	-RM1,978,620.00	49.3	14.921225	-0.504	3.3
KLCC	RM6,385,950.00	RM 4,507,802.00	-RM1,878,148.00	34.5	15.321321	-0.348	2.2
Bangsar	RM6,480,000.00	RM 4,920,957.00	-RM1,559,043.00	27.3	15.409014	-0.275	1.8
Ampang Hilir	RM5,800,000.00	RM 4,317,390.00	-RM1,482,610.00	29.3	15.278162	-0.295	1.9
Bangsar	RM5,115,000.00	RM 3,646,065.00	-RM1,468,935.00	33.5	15.109159	-0.339	2.2
Bangsar	RM6,600,000.00	RM 5,144,249.00	-RM1,455,751.00	24.8	15.45339	-0.249	1.6
Bangsar	RM6,660,000.00	RM 5,313,014.00	-RM1,346,986.00	22.5	15.48567	-0.226	1.4
Ampang Hilir	RM6,350,000.00	RM 5,081,078.00	-RM1,268,922.00	22.2	15.441034	-0.223	1.4
Bangsar	RM6,450,000.00	RM 5,203,855.00	-RM1,246,145.00	21.4	15.4649105	-0.215	1.4

Fig. 9. Sample of Dataset with High Value of Negative Percentage Difference.



Location	Price	SalePrice_PredXGB2	Difference	Percentage	SalePrice_PredXGE	Differen	Percentage
KLCC	RM2,200,000.00	RM 3,994,731.00	RM1,794,731.00	57.9	15.200487	0.597	4.0
KLCC	RM2,500,000.00	RM 3,940,634.00	RM1,440,634.00	44.7	15.186852	0.455	3.0
KLCC	RM1,680,000.00	RM 2,700,143.00	RM1,020,143.00	46.6	14.808816	0.475	3.3
Mont Kiara	RM1,800,000.00	RM 2,797,306.00	RM 997,306.00	43.4	14.844168	0.441	3.0
Bukit Tunku (Kenny Hills)	RM2,300,000.00	RM 3,277,969.00	RM 977,969.00	35.1	15.002735	0.354	2.4
KLCC	RM1,500,000.00	RM 2,439,551.00	RM 939,551.00	47.7	14.707325	0.486	3.4
Mont Kiara	RM1,300,000.00	RM 2,209,378.00	RM 909,378.00	51.8	14.608222	0.530	3.7
KLCC	RM2,700,000.00	RM 3,566,958.00	RM 866,958.00	27.7	15.087224	0.278	1.9
KLCC	RM2,700,000.00	RM 3,522,552.00	RM 822,552.00	26.4	15.074697	0.266	1.8
KLCC	RM2,413,375.00	RM 3,179,384.00	RM 766,009.00	27.4	14.9721985	0.276	1.9
KLCC	RM2,357,250.00	RM 3,117,411.00	RM 760,161.00	27.8	14.952514	0.280	1.9
KLCC	RM2,700,000.00	RM 3,457,221.00	RM 757,221.00	24.6	15.055976	0.247	1.7
Bangsar	RM2,700,000.00	RM 3,449,481.00	RM 749,481.00	24.4	15.053735	0.245	1.6
Mont Kiara	RM2,200,000.00	RM 2,937,095.00	RM 737,095.00	28.7	14.892932	0.289	2.0
Bukit Tunku (Kenny Hills)	RM2,200,000.00	RM 2,936,515.00	RM 736,515.00	28.7	14.892735	0.289	2.0
KLCC	RM2,357,250.00	RM 3,083,831.00	RM 726,581.00	26.7	14.941684	0.269	1.8

Fig. 10. Sample of Dataset with High Value of Positive Percentage Difference.

Fig. 9 and 10 illustrate that the highest negative and positive difference in predicted and actual price is also coming from a high-end location in Kuala Lumpur, which is KLCC. A group of high positive difference which is colored in darker green and a group of high negative difference which is colored in darker red is also coming from similar locations which are KLCC, Bangsar, Ampang Hilir, and Bukit Tunku. These locations are considered as high-end locations in Kuala Lumpur, thus the process of pricing a house might be difficult, resulting in the pricing of below or above market value.

VI. Result and Discussion

A comparison of all four models was conducted to determine which model delivers the most accurate results. The performance metrics used were mean absolute error (MAE), root mean squared error (RMSE) value, and adjusted R squared. The model that recorded the smallest MAE, RMSE value and adjusted R squared value closest to 1 was chosen and then used in the model deployment phase. Model comparison and evaluation results are shown in Table III. The model evaluation results comparing four models which are multiple linear regression, ridge regression, LightGBM, and XGBoost show that the XGBoost model has slightly better performance with the lowest MAE and RMSE values and has adjusted R-squared closest to one which indicated a good fit model compared to multiple linear regression, ridge regression, and LightGBM model. The XGBoost model was chosen as the best predictive model.

The aim of this study is to provide an accurate machine learning model for forecasting condominium house prices in Kuala Lumpur. A machine learning model has been proposed to evaluate the relationship between a dependent variable (housing price) and a series of independent variables (attributes). Multiple linear regression, ridge regression, LightGBM, and XGBoost were used and measured against each other. To evaluate the performance of the model, statistical measures such as mean absolute error and root mean square error was also established. The coefficient of determination (R-squared) was also derived to determine how accurately the model predicted the outcome. The XGBoost model was used in the deployment phase as this model was able to accurately predict house prices in Kuala Lumpur, with the highest coefficient of determination (R-squared), which means this model is best-fitted to the dataset. Even though this model was able to predict house prices with the best coefficient of determination value, however, there were still a high percentage difference in predicted house prices and actual house prices in several rows which is less than 5% of entire dataset. Based on these findings, it can be concluded that several locations which can be categorized as high-end locations such as Mont Kiara and KLCC were difficult to be priced. The price of the house might be below or above market value.

The proposed XGBoost model is the first application of XGBoost to the study of the Kuala Lumpur housing market. The model used in this analysis was able to tackle the problems of the housing market in Kuala Lumpur as the XGBoost model has a better fitting and predictive abilities. The XGBoost model was able to generate results that were more consistent and justifiable than other models used for housing market data. The XGBoost model achieved better predictive ability, with the lowest mean absolute error (MAE) and root mean squared error (RMSE), and adjusted R-squared value closest to 1, which indicates the most accurate model. In addition, consistent model performance was found in the XGBoost model as XGBoost outperformed other models in the training and testing R-squared value. The proposed XGBoost model is, therefore, effective in predicting housing prices, which favor not only future house buyers but also investors and policymakers in the real estate industry. In other words, the proposed model will be used to estimate the selling price of the house and then equate it with the currently offered price to know the actual market conditions.

TABLE III. Model Comparison and Evaluation Result

Model	MAE	MSE	RMSE	R-Sq*	R-Sq**	AdjR*
XGBoost*	0.148	0.039	0.197	0.921	0.912	0.911
LightGBM	0.161	0.044	0.210	0.902	0.899	0.898
MLR	0.181	0.057	0.238	0.872	0.871	0.869
RR	0.195	0.064	0.252	0.855	0.855	0.853

R-Sq* for Training Data; R-Sq** for Testing Data;

VII. CONCLUSION

This paper demonstrated the use of the advanced ML models on house price prediction based on the Kuala Lumpur housing data. The two most recent ML models, namely LightGBM and XGBoost were implemented and compared with the traditional models, namely Multiple Linear Regression and Ridge Regression. The results showed that the XGBoost model was the most promising with 0.0387 for the MSE and was used in the deployment phase. This model accurately predicts house prices in Kuala Lumpur, with the highest coefficient of determination (R-squared). Future works can include more attributes such as the size of the house and proximity to amenities, which can significantly affect the accuracy of the predicted outputs. In addition, future research might consider other locations in their study as this study is only focused on locations in the Kuala Lumpur region. Future research might expand the area of this research, whether to conduct and include the whole nation in Malaysia. This study can significantly assist in predicting of future house prices and the establishment of real estate policies.

ACKNOWLEDGMENT

The authors would like to thank the Faculty of Computer and Mathematical Sciences, and the Research Management Centre (RMC), Universiti Teknologi MARA, Malaysia, for the support throughout this research.

REFERENCES

[1] A. Soy Temür, M. Akgün, and G. Temür, "Predicting housing sales in turkey using arima, lstm and hybrid models," J. Bus. Econ. Manag., vol. 20, no. 5, pp. 920–938, 2019, doi: 10.3846/jbem.2019.10190.

[2] W. C. Choong, "Statistical Analysis Of Housing Prices In Petaling District Using Linear Functional Model." UTAR, 2018.

[3] M.H. Yop, "Harga Rumah Mengikut Negeri," 2018. https://www.data.gov.my/data/ms_MY/dataset/harga-rumah-mengikut-negeri.

[4] R. E. Febrita, A. N. Alfiyatin, H. Taufiq, and W. F. Mahmudy, "Data-driven fuzzy rule extraction for housing price prediction in Malang, East Java," in 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Oct. 2018, pp. 351–358, doi: 10.1109/ICACSIS.2017.8355058.

[5] S. Abdul-Rahman, N. F Kamal Arifin, M. Hanifiah and S. Mutalib, "Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier" International Journal of Advanced Computer Science and Applications(IJACSA), 12(9), 2021. http://dx.doi.org/10.14569/IJACSA.2021.0120950

[6] P. H. Damia Abd Samad, S. Mutalib, and S. Abdul-Rahman, "Analytics of stock market prices based on machine learning algorithms," Indones. J. Electr. Eng. Comput. Sci., vol. 16, no. 2, pp. 1050–1058, 2019, doi: 10.11591/ijeecs.v16.i2.pp1050-1058.

[7] N. S. Mohd Shafiee and S. Mutalib, "Prediction of Mental Health Problems among Higher Education Student Using Machine Learning," Int. J. Educ. Manag. Eng., vol. 10, no. 6, pp. 1–9, 2020, doi: 10.5815/ijeme.2020.06.01.

[8] N. Mohammad Suhaimi, S. Abdul-Rahman, S. Mutalib, N. H. Abdul Hamid, and A. Md Ab Malik, Predictive Model of Graduate-On-Time Using Machine Learning Algorithms, vol. 1100, no. September. Springer Singapore, 2019.

[9] Y. F. Chang, W. C. Choong, S. Y. Looi, W. Y. Pan, and H. L. Goh, "Analysis of housing prices in Petaling District, Malaysia using functional relationship model," Int. J. Hous. Mark. Anal., vol. 12, no. 5, pp. 884–905, Oct. 2019, doi: 10.1108/IJHMA-12-2018-0099.

[10] J. Manasa, R. Gupta, and N. S. Narahari, "Machine Learning based Predicting House Prices using Regression Techniques," in 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Mar. 2020, pp. 624–630, doi: 10.1109/ICIMIA48430.2020.9074952.

[11] M. Thamarai and S. P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning," Int. J. Inf. Eng. Electron. Bus., vol. 12, no. 2, pp. 15–20, 2020, doi: 10.5815/ijieeb.2020.02.03.

[12] H. Wu et al., "Influence factors and regression model of urban housing prices based on internet open access data," Sustain., vol. 10, no. 5, pp. 1–17, 2018, doi: 10.3390/su10051676.

[13] J. B. H. Yap and X. H. Ng, "Housing affordability in Malaysia: perception, price range, influencing factors and policies," Int. J. Hous. Mark. Anal., vol. 11, no. 3, pp. 476–497, Jun. 2018, doi: 10.1108/IJHMA-08-2017-0069.

[14] A. Abdullahi, Usman, H. Usman & I. Ibrahim, "Determining house price for mass appraisal using multiple regression analysis modeling in Kaduna North, Nigeria," ATBU Journal of Environmental Technology, 11(1), 26-40, 2018.

[15] T. Mohd, S. Masrom, and N. Johari, "Machine learning housing price prediction in petaling jaya, Selangor, Malaysia," Int. J. Recent Technol. Eng., vol. 8, no. 2 Special Issue 11, pp. 542–546, 2019, doi: 10.35940/ijrte.B1084.0982S1119.

[16] A. Jafari and R. Akhavian, "Driving forces for the US residential housing price: a predictive analysis," Built Environ. Proj. Asset Manag., vol. 9, no. 4, pp. 515–529, Sep. 2019, doi: 10.1108/BEPAM-07-2018-0100.

[17] A. Nur, R. Ema, H. Taufiq, and W. Firdaus, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study : Malang, East Java, Indonesia," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 10, pp. 323–326, 2017, doi: 10.14569/ijacsa.2017.081042.

[18] J. H. Chen, C. F. Ong, L. Zheng, and S. C. Hsu, "Forecasting spatial dynamics of the housing market using Support Vector Machine," Int. J. Strateg. Prop. Manag., vol. 21, no. 3, pp. 273–283, 2017, doi: 10.3846/1648715X.2016.1259190.

[19] Pai, P.F. and Wang, W.C., "Using machine learning models and actual transaction data for predicting real estate prices". Applied Sciences, 10(17), p.5832, 2020.

[20] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," in 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Dec. 2018, pp. 35–42, doi: 10.1109/iCMLDE.2018.00017.

[21] Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing Price Prediction via Improved Machine Learning Techniques," Procedia Comput. Sci., vol. 174, no. 2019, pp. 433–442, 2020, doi: 10.1016/j.procs.2020.06.111.

[22] K. Zhang, L. Shen, and N. Liu, "House Rent Prediction Based on Joint Model," in Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition, Oct. 2019, pp. 507–511, doi: 10.1145/3373509.3373578.

[23] Y. Zhou, "Housing Sale Price Prediction Using Machine Learning Algorithms," 2020, [Online]. Available: https://escholarship.org/uc/item/0th2s0ss.

[24] P. Katherina, "Data and Methodology", How to Write About Economics and Public Policy, pp 241-270, 2018.

[25] Watkins, M. W. A Step-by-Step Guide to Exploratory Factor Analysis with SPSS. Routledge, 2021

[26] Farmer, Antoinette Y., Antoinette Y. Farmer, and G. Lawrence Farmer. Research methods for social work: A problem-based approach. SAGE Publications, 2020.

[27] J. W. Osborne and E. Waters, "Four assumptions of multiple regression that researchers should always test," Pract. Assessment, Res. Eval., vol. 8, no. 2, pp. 2002–2003, 2002.

[28] N. E. Jeremia, S. Nurrohmah and I. Fithriani, "Robust Ridge regression to solve a multicollinearity and outlier", J. Phys.: Conf. Ser. 2020

[29] N. Deepa et al., "An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier," J. Supercomput., vol. 77, no. 2, pp. 1998–2017, Feb. 2021, doi: 10.1007/s11227-020-03347-2.

[30] W. Xu, X. Liu, F. Leng, and W. Li, “Blood-based multi-tissue gene expression inference with Bayesian ridge regression,” Bioinformatics, vol. 36, no. 12, pp. 3788–3794, Jun. 2020, doi: 10.1093/bioinformatics/btaa239.

[31] Y. Yang and Y. Yang, “Hybrid prediction method for wind speed combining ensemble empirical mode decomposition and bayesian ridge regression,” IEEE Access, vol. 8, pp. 71206–71218, 2020, doi: 10.1109/ACCESS.2020.2984020.

[32] C. Fan, Z. Cui, and X. Zhong, “House Prices Prediction with Machine Learning Algorithms,” in Proceedings of the 2018 10th International Conference on Machine Learning and Computing, Feb. 2018, pp. 6–10, doi: 10.1145/3195106.3195133.

[33] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, “A hybrid regression technique for house prices prediction,” IEEE Int. Conf. Ind. Eng. Eng. Manag., vol. 2017-December, no. August 2018, pp. 319–323, 2018, doi: 10.1109/IEEM.2017.8289904.

[34] L. Zhu and L. Li, “Comparison of Regression Models on House Value Prediction,” 2020.

[35] H. Zeng et al., “A LightGBM-Based EEG Analysis Method for Driver Mental States Classification,” Comput. Intell. Neurosci., vol. 2019, 2019, doi: 10.1155/2019/3761203.

[36] C. Chen, Q. Zhang, Q. Ma, and B. Yu, “LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion,” Chemom. Intell. Lab. Syst., vol. 191, no. June, pp. 54–64, 2019, doi: 10.1016/j.chemolab.2019.06.003.

[37] Y. Song et al., “Prediction of Double-High Biochemical Indicators Based on LightGBM and XGBoost,” in Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, Jul. 2019, pp. 189–193, doi: 10.1145/3349341.3349400.

[38] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[39] J. Eronen, “Housing unit price prediction system.” 2018.

[40] N. H. Zulkifley, S. A. Rahman, N. H. Ubaidullah, and I. Ibrahim, “House price prediction using a machine learning model: A survey of literature,” Int. J. Mod. Educ. Comput. Sci., vol. 12, no. 6, pp. 46–54, 2020, doi: 10.5815/ijmecs.2020.06.04.

[41] J. J. Wang et al., “Predicting House Price with a Memristor-Based Artificial Neural Network,” IEEE Access, vol. 6, pp. 16523–16528, 2018, doi: 10.1109/ACCESS.2018.2814065.

[42] M. F. Mukhlishin, R. Saputra, and A. Wibowo, “Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor,” in 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), Nov. 2017, pp. 171–176, doi: 10.1109/ICICOS.2017.8276357.

[43] A. Varma, A. Sarma, S. Doshi, and R. Nair, “House Price Prediction Using Machine Learning and Neural Networks,” in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Apr. 2018, pp. 1936–1939, doi: 10.1109/ICICCT.2018.8473231.

[44] R. Reed, “The relationship between house prices and demographic variables,” Int. J. Hous. Mark. Anal., vol. 9, no. 4, pp. 520–537, Oct. 2016, doi: 10.1108/IJHMA-02-2016-0013.

[45] A. Kallner, “Formulas,” in Laboratory Statistics, A. B. T.-L. S. (Second E. Kallner, Ed. Elsevier, 2018, pp. 1–140.

[46] J. F. Hair, W. C. Black, B. J. Babin and R. E. Anderson (2019). Multivariate data analysis, 2019.

[47] Byrne, B.M. (2016). Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming, Third Edition (3rd ed.). Routledge. https://doi.org/10.4324/9781315757421, 2016.

[48] G. S. K. Ranjan, Verma, A. K. and R. Sudha. "K-nearest neighbors and grid search cv based real time fault monitoring system for industries." In 2019 IEEE 5th international conference for convergence in technology (I2CT), pp. 1-5. IEEE, 2019.

[49] J. Macaluso, “Testing Linear Regression Assumptions in Python.,” 2018. https://jeffmacaluso.github.io/post/LinearRegressionAssumptions/.