Using Machine Learning to Forecast Residential Property Prices in Overcoming the Property Overhang Issue

Lim Wan Yee
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia
wanyee.lim@graduate.utm.my

Nur Azaliah Abu Bakar
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia
azaliah@utm.my

Noor Hafizah Hassan
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia
noorhafizah.kl@utm.my

Norziha Megat Mohd Zainuddin
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia
norziha.kl@utm.my

Rasimah Che Mohd Yusoff
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia
rasimah.kl@utm.my

Nor Zairah Ab Rahim
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia
nzairah@utm.my

***Abstract*— Overhang property issue has sustained over the past ten years in Malaysia. Major overhang property issue was contributed from the unsold residential property. Though the government announced to build a data system and provide the housing data to prevent a mismatch of supply-demand in the property market, there are still not many relevant studies or research on predicting residential property prices. Hence, it is essential to understand the factors that influence the price of residential properties. The study aims to predict the price of a residential property by using a machine learning algorithm. Three algorithms were selected, namely Decision Tree, Linear Regression, and Random Forest, tested against the training and testing datasets obtained from the Malaysian Valuation and Property Services Department. Results show that the Random Forest model produced high accuracy with lower r_squared (R^2), RMSE, and MAE values. Significantly, the study has contributed a new insight into essential property features that primarily influence the property price, which will be useful for property developers and buyers who wish to invest in the property market.**

Keywords—Decision Tree, Linear Regression, Machine Learning algorithms, Random Forest, property price

I. INTRODUCTION

Unsold property is undoubtedly a long-standing problem that happened in Malaysia. Over the past decade, the government has actively made and implemented a housing schema to reduce the slow sales of property. However, after so many years, the problem persists and becomes more severe over the year. Besides, the epidemic of Covid-19 occurred in December 2019, and the government has announced a Movement Control Order (MCO), and eventually, many people lost their jobs [1]. This situation is inevitable to make the overhang property issue worse [2]. The number of unsold residential properties climbing year by year due to the mismatch between the affordability of houses built and offered in the market. It can be seen from the housing financing schemas, "Home Ownership Campaign (HOC)", which is the proposed incentives did not simulate enough demand and are not likely to push overall sales of the property effectively or alleviate the persisting overhang issue [3]. Aside from that, the government also announced to build a data system and provide the housing data to the relevant party to prevent supply-demand mismatch in the property market [4]. However, there are still no relevant studies or research on predicting sell-through rates in a residential unit.

Some industry players and researchers believed that the absence of accurate data for the developer would ensure the supply-demand mismatch [5-7]. Without the correct data, developers could not provide suitable properties based on the price and location. As a result, it exacerbates the oversupply issue of property. This massive overhang property does not create any rolling economic effect from the economy's perspective while impacting nearly 150 industries relevant to the real estate sector [9].

Although the government has made an effort to build up the data system in three years, there is still a lack of studies related to predicting housing price, making this study unable to compare with other related work [7]. These undoubtedly would cause a lag and continue to weigh down economic growth in Malaysia. As a result, the overhang of the unfulfilled property issue urgently needs to be resolved. The real estate industry is one of the most critical aspects of every country's economy. Real estate buyers and sellers and economists benefit from observations of the real estate market and correct projections of real estate values. However, real estate forecasting is a sophisticated and challenging process due to numerous direct and indirect elements that unavoidably impact prediction accuracy [8].

Therefore, this study's motivation is to explore factors that influence residential properties' prices and predict the suitable price for the developer. The targeted goal will be accomplished by identifying the variables that affect residential properties' prices and then applying a machine-learning algorithm to forecast the residential property's price. The result will be evaluated based on the accuracy of the machine learning algorithm proposed.

II. LITERATURE REVIEW

Owning a roof over our head often imparts a sense of stability and belonging, unlike renting a house brings the possibility of eviction. For that reason, people are more

2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET) | 978-1-6654-2899-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IICAIET51634.2021.9573830

978-1-6654-2899-6/21/$31.00 ©2021 IEEE

inclined to buy a house if the economic condition allows [2]. Over the year from 2010 to 2014, Johor, Selangor, and Sabah have always had the highest unsold property compared to other states. On the other hand, Kelantan, Terengganu, Perlis, and Labuan have a relatively low unsold unit. In contrast, WP Putrajaya has no new launch residential property, making it has no overhang property [9]. From StarProperty "The 2019 Buyer Sentiment Survey" conducted for 11 months, with more than 5,000 respondents, 86% of respondents preferred to buy a house [1]. Nevertheless, Johor state was considered the most severe overhang residential issue with several unsold units since 2010 and hit 13,000 units at Quarter 2 in 2019 [10]. Based on this situation, it is interesting to explore why property overhang happened in Johor compared to other places in Malaysia.

Despite understanding economic and social factors as highlighted from studies by [2, 10], venturing into data analytics as the solution will bring new insight into this rising issue. The Machine Learning (ML) algorithm consists of self-learning characteristics by learning the pattern and trends of the dataset. Therefore, ML can potentially resolve property overhang since it can deal with large and complicated datasets like residential property sales [5]. To accurately predict the residential property price, the ML features should be considered as much as possible. A study by [5] provides a novel technique that demonstrates how machine learning algorithms may forecast home values utilising standard housing characteristics as input. The authors explain the usage of Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM) to investigate a data sample of around 40,000 housing transactions in Hong Kong over more than 18 years. It was discovered that SVM provides relatively accurate predictions under time constraints.

Meanwhile, [8] employed quantitative data obtained from actual transaction data documenting specifics of real estate transaction data in Taiwan rather than qualitative variables impacting real estate pricing. In this work, empirical data indicated that attribute selection for machine learning models improves predicting accuracy of four forecasting models. This study used four machine models with feature selection to forecast real estate transaction prices in Taichung, Taiwan, using actual transaction data. These models included Least Squares Support Vector Regression (LS-SVR), Classification and Regression Tree (CART), General Regression Neural Networks (GRNN) and Back-propagation Neural Networks (BPNN). Their findings indicate that the LS-SVR model outperformed the other three forecasting models in terms of prediction accuracy.

Alternatively, [10] use regression techniques such as Multiple Linear Regression (MLR), Ridge and Lasso Regression, Support Vector Regression (SVR), and boosting algorithms such as Extreme Gradient Boosting (XGBoost) to develop a predictive model for determining the sale price of houses in Bengaluru cities. The housing price prediction based on Convolutional Neural Network (CNN) is performed by [11] for house price forecast based on a dataset from the Land Resources and Housing Information Center for Dalian, China, including 16 characteristics. Besides, multiple ML algorithms are used in research by [12] to forecast housing prices in Petaling Jaya, Selangor, Malaysia, based on attributes such as several beds and floor level building age and floor size. Their research analyses the outcomes of RF, Decision Tree, (DT) Ridge Regression (RR), Linear Regression (LR), and LASSO. The results conclude that RF is the best overall accuracy as measured by the Root Mean Squared Error (RMSE).

A study by [13] perform a housing price prediction based on deep learning by comparing the best ML algorithm between CNN, Long Short Term-Memory (LSTM) and Autoregressive–Moving-Average (ARMA). They found out that CNN performs better even with all irrelevant attributes. Meanwhile, LSTM performs better when added with the time factor. ARMA only highlighted time series, ignoring other attributes causing high prediction error. On the other hand, [14] examined housing data from 5,359 townhouses in Fairfax County, Virginia, between 2004 and 2007, consisting of 76 characteristics, of which 28 were finally selected after filtering using t-test and logistic regression. Sixteen of these elements are the number of bedrooms and bathrooms, number of fireplaces, overall space size, number of cooling and heating systems, and type of parking. Most findings stated that LR handles well in a regression problem, but performance will decrease when variables increase. The RF better captures the nonlinear relationship between building features and the selling price, providing a better estimation. Hence, RF is utilised because it also reduces data overfitting [5]. The previous study has also shown that RF is a reliable algorithm that predicts property prices accurately [12].

Meanwhile, DT uses a tree structure to describe the relationships in a database to trace a route in a tree to categorise a new instance [14]. It is a combination of regression and classification model that constructed in tree-structure. A DT will break down the data set into smaller subsets, thus creating and growing the decision tree simultaneously, and the output is in the tree structure form consisting of decision numbers and leaf nodes. The node of decision denotes the indicator which has few branches. In contrast, the node of lead represents a decision of classification. The highest decision node in the DT is called the root node, which matches the best predictor [15]. Hence we can use it to find patterns in any property price data by learning decision rules. On the other hand, LR is the study of two independent variables to establish a single connection, and it is a valuable metric for technical and quantitative research in financial markets [10]. LR can identify the variables affecting house prices in predicting property prices, such as area, number of rooms, bathrooms, etc. Considering the previous studies' considerations, these three machine learning algorithms (Decision Tree, Linear Regression, and Random Forest) will be selected to provide the most accurate ML classifiers in this study context.

III. Methodology

The section presents the study's process flow, divided into four main phases: data collection and preparation, data analysis, model development, and evaluation and implementation and conclusion. Fig. 1 depicts the overall process flow.

A. Data Collection

In this research, the residential property dataset is collected from open-source data on Brickz (https://www.brickz.my/). The source of data on Brickz is from the Valuation and Property Services Department or Jabatan Penilaian and Perkhidmatan Harta (JPPH) Malaysia. Brickz provides property information from all states in Malaysia that had paid for the Sales and Purchase Agreement (SPA). This data source captured all the official recorded transactions since January 2014, and the data is updated monthly. For this study, the dataset included 2,087 property transactions from April 2017 to Dec 2019 and was randomly selected from Brickz. Apart from collecting valid data, this raw dataset also consists of null values, noise variables, outliers, and unrelated variables that subsequently affect the algorithm's learning process. Thus, before entering the subsequent phases, we need to identify the dataset according to the data types and the abnormal data's actions.

B. *Data Preparation*

This study collected 44 datasets from different areas in Johor with a various number of variables. For example, the building type of condominium and flat has only eight variables, while the other building type consists of ten variables. Integrating all 44 areas of datasets would require the same number of variables, same variable names, and the same data types for each variable. Therefore, to achieve this, each dataset will undergo pre-processing to ensure all the datasets standardise on the features for data integration. The dataset is then presented in Excel format with a separate sheet of each area. This is an ease of view of the number of variables lacking for the dataset. The most significant number of variables which is ten, will be the benchmarks for the other datasets. The dataset with less than ten variables will be added, respectively, as null values.

After all the 44 sets of data have been pre-processed to common features, they are integrated into a single dataset based on the variables. The integration process aims to ease data analysis, natural language processing, and data training. This process is done using Microsoft Query in Excel. The integrated dataset consists of 2,087 transactions and ten variables, with three qualitative and seven quantitative data. Data preparation for individual datasets will standardise all the data and integrated them into a single dataset. However, the integrated dataset might contain null values, outliers, and data types that are unable to be trained by algorithms. Thus, it needs human input programming language that is precise and highly structured. So this study applies Natural Language Processing (NLP) using Jupyter Notebook to transform the ordinal data into quantitative data to let the machine understand the text.

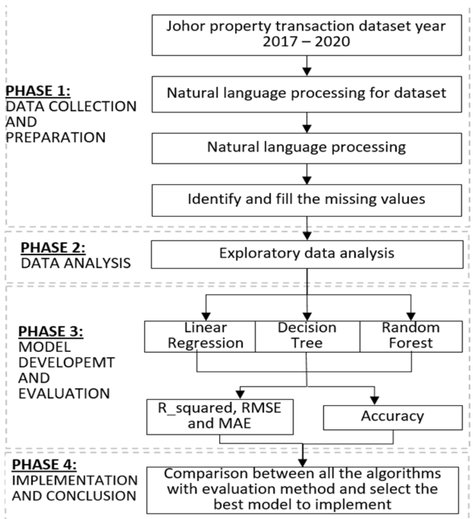

Fig. 1. Process Flow of the Research.

C. *Data Analysis*

In this phase, the exploratory data analysis is conducted to provide an insight into the character of the dataset in terms of mean, median, mode, standard deviation, and minimum and maximum values. There are 14 variables with three nominal data in the dataset: address, building type, and tenure. Meanwhile, quantitative data consist of SPA date, floors, rooms, land area, built up, price psf, and price. Table I shows the variables after data preparation and analysis.

D. *Model Development and Evaluation*

The dataset now is cleaned from the null value, with the value and information are now standardised to one format for every respective variable. Thus, data cleaning is not required anymore and proceed with the modelling process. The dataset is separated into a training set and a testing set. The training set will take up 80% of the data, which amounts to 1,670 property transactions and, 20% of the data which 417 property transactions will be contributed to the testing set. The three selected machine learning algorithms, Decision Tree (DT), Linear Regression (LR) and Random Forest (RF), are now ready for the data training process. This study uses Jupyter Notebook to assist in modelling construction. The default parameters for all machine learning algorithms used in this dataset. After the training process is completed, the ML algorithms' outcomes are tested with the testing set and evaluated based on the several metrics shown in Table II. The model with the best performance in terms of higher accuracy and R-Squared Value (R^2) with lower Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) will be chosen through the evaluation process conducted. In addition, the selection of the variables before training data is essential to prevent the weak relationship variables from affecting the data training process, resulting in inaccurate prediction outcomes. Fig. 2 shows the correlation matrix, and the correlation coefficient is compared between every variable in this dataset.

TABLE I. LIST OF VARIABLES IN PROPERTY DATASET AFTER DATA PREPARATION AND ANALYSIS

No	Variable	Description
1	SPA Date	Date when Sales and purchase agreement (SPA) has been signed by both seller and buyer parties. There is no further negotiation on the property once signed
2	Address	The address of the residential property contains the street name and area
3	Building Type	The building contains (Apartment, Bungalow, Cluster House, Cluster house – intermediate, Condominium, Flat, Semi-D, Service residence, Terrace house - corner lot, Terrace house – end lot, Terrace house - intermediate and Townhouse)
4	Tenure	Condition of land is Freehold or Leasehold
5	Floors	Number of floors of a building
6	Rooms	Number of rooms of a building
7	Land Area	The total area of the building sits on it
8	Built-Up	The total area of the building being constructed
9	Price PSF	Price per square feet of a building
10	Price	The purchase price of the building started on SPA
11	Month	Month extracted from SPA data
12	Area	Area of the building located, in total 44 areas
13	Tenure_f	Freehold: 1; Leasehold: 2
14	Building Type_f	Apartment: 1, Bungalow:2, Cluster house:3, Cluster house – intermediate:4, Condominium:5, Flat:6, Semi-D:7, Service residence:8, Terrace house - corner lot:9, Terrace house – end lot:10, Terrace house - intermediate:11 and Town house:12)

TABLE II. EVALUATION METRICS USED

Evaluation metric	Expected value	Details
Accuracy	High	Accuracy indicates the percentage of values to be predicted correctly
R^2	High	Showing the correlation between variables and the target variable. Higher the R^2, the more correlated to the fitted line
RMSE	Low	Lower the RMSE indicates the square root of distance between the expected value and predicted value are closer and more accurate
MAE	Low	Lower the MAE indicates the average distance of all data between the expected value and the predicted value is close and more accurate

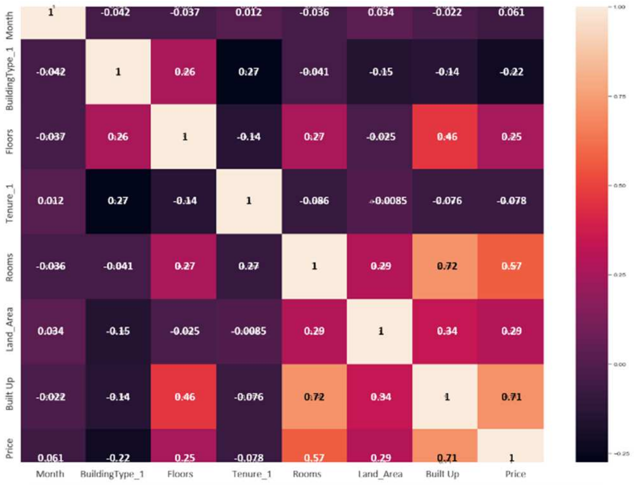

Fig. 2. Correlation Matrix of a Dataset.

Most of the relationship between the two variables is relatively moderate. Variables that are weakly or not related are considered as noise in the learning process of the machine. "Rooms" and "Built-up" are moderately correlated to "Price", while "Tenure" and "Month" are weakly correlated to the target variable.

IV. RESULTS AND DISCUSSIONS

This section evaluates the DT, LR and RF models based on accuracy, R^2, RMSE and MAE values. This section is separated into two subsections for before and after removing the weakly correlated variables.

A. *Decision Tree Results*

The first result is using DT, as shown in Fig. 3. It explains the frequency distribution of actual and predicted prices by DT trained with all variables. The predicted values of price are represented in orange, while actual values are represented in blue.

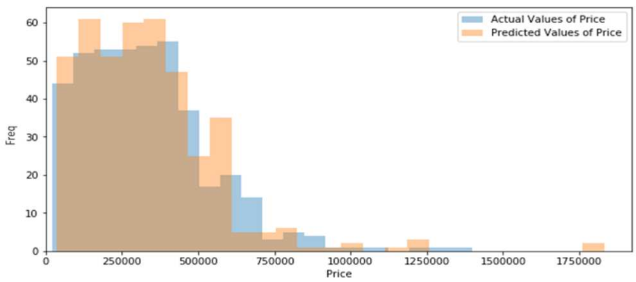

Fig. 3. Combination of a frequency distribution of actual and predicted prices by DT trained with all variables.

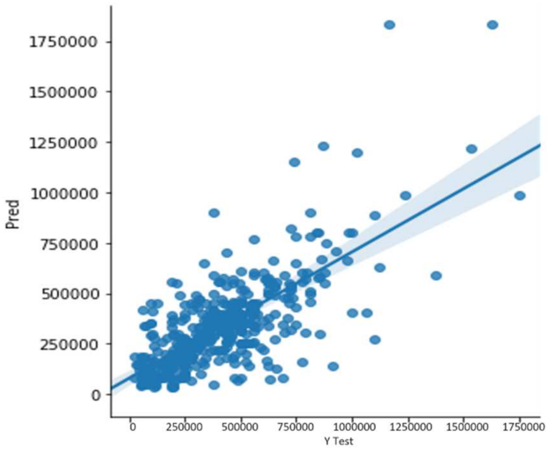

Fig. 4. Scatter plot of predicted prices by DT trained with all variables.

Both of the values are overlapping to each other for ease of comparison. It can be seen that most of the areas are overlapping between actual and predicted values. The predicted value is achieved at 99.78% accuracy. The R^2 value is 0.39, RMSE is RM164,472.02 and MAE is RM110,142.72. Although the accuracy is relatively high, yet the R^2 value is relatively low. R^2 is a method used to measure how close the data are fitted to the regression line. The closer the data fitted to the regression line, the higher the R^2. In this algorithm, the R^2 is considered relatively low, where all the data are dispersed around the regression line, as shown in Fig. 4.

The accuracy of prediction values in this dataset are lower than then dataset without removing weakly correlated variable. However, this isn't indicating the dataset is worse than the previous dataset. RMSE and MAE will require to be considered in the evaluation. After the weakly correlated variables have been removed, the magnitude of the error has decreased. The predicted values are more fitted to the regression line with R^2 of 0.41, 162,090.62 RMSE, and 99,408.18 of MAE. Table III shows the differences discussed.

B. *Linear Regression Result*

Next is LR results, as shown in Fig. 5. It depicts the frequency distribution of actual and predicted prices by LR trained with all variables.

TABLE III. DECISION TREE ALGORITHM OUTCOMES

Dataset	Accuracy(%)	R^2	RMSE(RM)	MAE(RM)
All variables	99.78%	0.39	164,472.02	110,142.72
Removed variables	98.62%	0.41	162,090.62	99,408.18
Difference	1.16	-0.02	2381.40	10,734.54

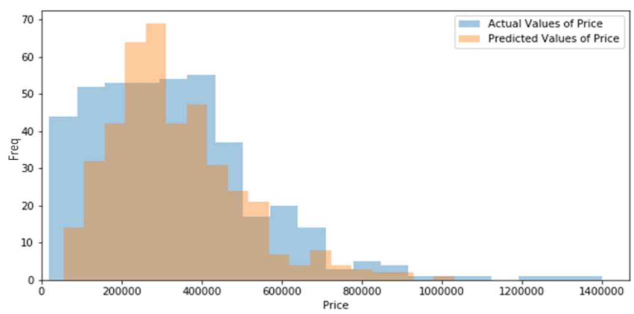

Fig. 5. Combination of the frequency distribution of actual and predicted prices by LR trained with all variables.

The overlapping area between the actual and predicted value is relatively low. The predicted price values have only 53.5% are predicted same values of price with the actual value. It shows that the predicted values of prices have enormous differences for a residential property below the price of RM500,000. The actual values indicating is flattened normal distribution graph while the algorithm tends to predict the values in a normal distribution curve, as shown in Fig. 6.

The R^2 is 0.56 and moderately correlated with the target variable. The RMSE and MAE have a high magnitude of error, showing 140,427.34 and 105,398.61 average differences from the fitted regression line. The accuracy decreased to 52.84% after removed the weakly correlated variables. The overlapping area is reduced compared to removing the variables. Multicollinearity phenomenon is often happening in the regression problem. The R^2 value is 0.55, which is also moderately correlated to the targeted variable. The RMSE is 140,961.27, and MAE is 105,645.06. The differences in the magnitude of error are increased to RM533.93 and RM246.45. The scatter plot shows the data are relatively sparse and far from the regression line. This is shown in Table IV.

C. Random Forest Result

Fig. 7 depicts the frequency distribution of actual and predicted prices by RF trained with all variables. The overlapping area between the predicted and the actual values are relatively high.

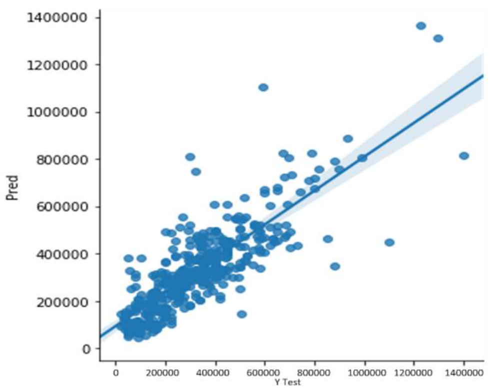

Fig. 6. Scatter plot of predicted prices by LR trained with all variables.

TABLE IV. LINEAR REGRESSION ALGORITHM OUTCOMES

Dataset	Accuracy(%)	R^2	RMSE(RM)	MAE(RM)
All variables	53.50%	0.56	140,427.34	105,398.61
Removed variables	52.84%	0.55	140,961.27	105,645.06
Difference	0.66	0.01	-533.93	-246.45

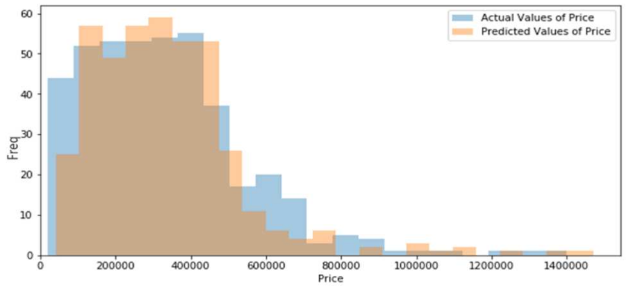

Fig. 7. Combination of the frequency distribution of actual and predicted prices by RF trained with all variables.

The accuracy of the predicted result was achieved at 93.58%. It is interesting to see that, although the accuracy is lower compared with the decision tree algorithm. Yet, the distribution pattern tends to follow the pattern of the actual values, as shown in Fig. 8. The R^2 is 0.69, which is moderately high correlated with the targeted variable. The RMSE and MAE are 117,318.45 and 83,679.72, respectively. From the scatter plot may observe the data tends to be plotted closer along the regression line. The accuracy of the predicted values of prices has decreased to 92.86%. From the pattern of the predicted values of price, the predicted result went huge differences for residential property around RM300,000. The R^2 has the same value before removing the variables, and the RMSE shows RM117,380.03 while MAE is 79,701.47. The RMSE has increased, saying the average magnitude of error has increased, while the magnitude of error for single data has decreased RM 3,978.25. Table V shows the results with all variables and removed variables.

D. Selection for Machine Learning Algorithm

This study compared and evaluated three algorithms' performance: DT, LR and RF. The ML algorithm that produced the highest accuracy is DT, which trained with all variables dataset, with 99.78% accuracy. The highest R^2 value among these algorithms is 0.69, created by an RF algorithm trained by all variables and removed variables dataset. The ML algorithm that generated the lowest values of RMSE and MAE is an RF which is RM117,318.45 and RM79,701.47. Besides the metric measurement, the predicted pattern is also essential to consider in the evaluation process. They are comparing the top two highest accuracy algorithms, which are decision tree and random forest. Both algorithms have performed well in terms of accuracy.

However, when observing the predicted pattern in all residential property prices, the RF gives a more exact pattern and more minor differences with actual values. At the same time, the DT has high accuracy when predicted on the lower and higher range of residential property, though, and the predicted values have many differences in the middle range house. Additionally, it can be observed that all three algorithms have meet difficulties in predicting the price range of around RM300,000.

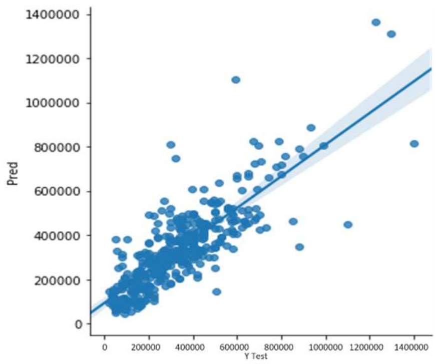

Fig. 8. Scatter plot of predicted prices by RF trained with all variables.

TABLE V. RANDOM FOREST ALGORITHM OUTCOMES

Dataset	Accuracy(%)	R^2	RMSE(RM)	MAE(RM)
All variables	93.58%	0.69	117,318.45	83,679.72
Removed variables	92.86%	0.69	117,380.03	79,701.47
Difference	0.72	0	-61.58	3,978.25

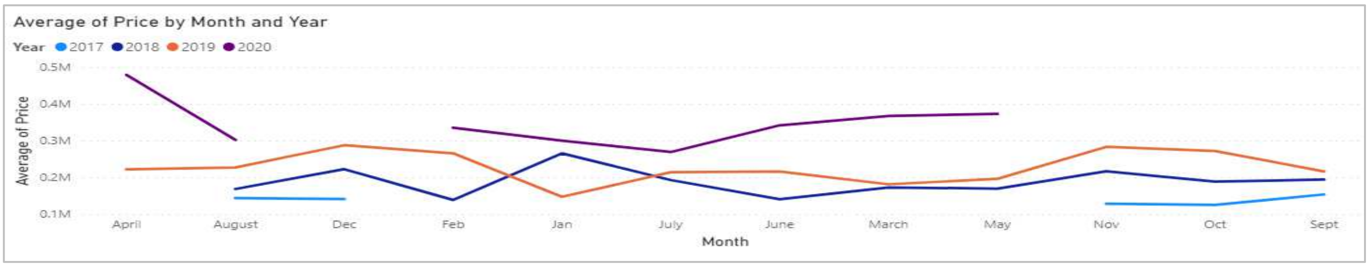

Fig. 9. Collected data plotted in time series.

One of the reasons causing this is due to almost all types of buildings having overlapping in this price range. The ability to predict the middle price range of residential property is essential. The average price range of residential property in Johor falls around RM300,000. To have most residential property predicted correctly, the algorithm must learn the characteristics of different buildings in this price range. Unlike building type "bungalow", the price for bungalow is higher than other building types or the land area and built up has much more than the other building type, making the algorithm more easily predict on it. For example by inputting such data as "Month: 3", "BuildingType:11; Terrace house-intermediate", "Floors: 1", "Tenure_f:1; Freehold", "Rooms: 3", "Land area: 1320 square feet" and "Built-up: 860 square feet", the output result is RM260,400. The actual residential price sold at RM250,000, which has a 4% difference. However, due to the incomplete data obtained, the data plotted in the time-series graph in Fig. 9 shows discontinuity for the dataset for 2017 and 2020. Nonetheless, the RF has predicted the best among other algorithms.

V. Conclusion

This study aims to construct a machine learning algorithm that allows predicting the residential property price in Johor. By comparing three machine learning algorithms in terms of evaluation based on the accuracy, R^2, RMSE and MEA. This model allows property developers to list better prices to their targeted customers and allot buyers to investigate market prices of similar building features. Interestingly, using ML to resolve this issue has brought out a new dimension of research. We explore the ability to handle the missing values and study the fundamental dataset characteristic from the property industry. In addition, this study is also exploring the potential use of Python in Jupyter Notebook from data preparation, data analysis, model development, model evaluation, and implementation. Nevertheless, we face several limitations and challenges throughout the project, such as the limited number of free and open datasets. It requires to pay up to RM450 for the yearly subscription to access the complete residential property dataset. In addition, there is a variant of the dataset provided. The dataset will be less variable, and it might be captured during different time ranges by other areas. This is closely related to the issue of specific price ranges of the residential property consist of too many different variables, which result from its lower accuracy for these price ranges of residential property.

However, performance variations in real estate may be attributed to various factors, including country, culture, market trends, and economic situations. Additionally, buyer and owner expectations are constantly changing due to changes in lifestyles, construction materials, environmental legislation, and rules requiring energy efficiency, as indicated by past studies by [8, 16]. Therefore, future research should seek to address this issue by increasing the diversity of the dataset by including other features like the residential geographical area, facilities, criminal rates, number of years and conditions of the house, and the average income of the residents. Furthermore, the modelling process can be improved by considering additional steps to cater to imbalanced data during the data preparation process. In summary, from this study, we can explore that the overhang issue has existed in Malaysia for an extended period, and the impact of the overhang property affects the developer.

References

[1] C. Viktor, "More Malaysians Willing To Buy A Home In 2020," in *StarProperty*, ed, 2020.

[2] N. Ab Rahim, S. Adzhar, N. Basrah, R. Ab Majid, and S. Mustafar, "Factors lead to overhang in affordable housing: A content analysis in qualitative approach," in *IOP Conference Series: Earth and Environmental Science*, 2019, vol. 385, no. 1, p. 012073: IOP Publishing.

[3] S. Ismail *et al.*, "Rethinking housing: between state, market and society: a special report for the formulation of the national housing policy (2018–2025), Malaysia," *Kuala Lumpur: Khazanah Research Institute,* 2019.

[4] J. IQI, "Property Survey and Index Malaysia Q1 2020 " in "Juwai IQI," Juwai IQI2020, Available: http://downloads.juwai.com/pdf/Property-Survey-and-Index-Malaysia-Q1-2020.pdf.

[5] W. K. Ho, B.-S. Tang, and S. W. Wong, "Predicting property prices with machine learning algorithms," *Journal of Property Research,* vol. 38, no. 1, pp. 48-70, 2021.

[6] A. Saull, A. Baum, and F. Braesemann, "Can digital technologies speed up real estate transactions? " *Journal of Property Investment & Finance,* 2020.

[7] N. A. A. Bakar and A. Yaacob, "The Role of Service Digitalisation in Realty Business during Pandemic from the Perspective of Real Estate Agency in Klang Valley, Malaysia," *International Journal of Real Estate Studies (INTREST),* vol. 14, no. 1, pp. 1-8, 2020.

[8] P.-F. Pai and W.-C. Wang, "Using machine learning models and actual transaction data for predicting real estate prices," *Applied Sciences,* vol. 10, no. 17, p. 5832, 2020.

[9] S. Ng, "What shall we do with the property overhang?," in *EdgeProp.my*, ed, 2020.

[10] J. Manasa, R. Gupta, and N. Narahari, "Machine learning-based predicting house prices using regression techniques," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2020, pp. 624-630: IEEE.

[11] Y. Piao, A. Chen, and Z. Shang, "Housing price prediction based on CNN," in *2019 9th International Conference on Information Science and Technology (ICIST)*, 2019, pp. 491-495: IEEE.

[12] 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, pp. 542-546, 2019.

[13] L. Yu, C. Jiao, H. Xin, Y. Wang, and K. Wang, "Prediction on housing price based on deep learning," *International Journal of Computer and Information Engineering,* vol. 12, no. 2, pp. 90-99, 2018.

[14] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert systems with applications,* vol. 42, no. 6, pp. 2928-2934, 2015.

[15] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets,* pp. 1-11, 2021.

[16] A. Minnullina, "Residential real estate value assessment considering the influence of various factors," in *E3S Web of Conferences*, 2019, vol. 91, p. 08049: EDP Scienc