Prediction of House Prices using Machine Learning Techniques

ABSTRACT

It was reported that 73 percent of unsold residential properties were priced beyond the means of average Malaysians to purchase them. The state of Johor had the biggest number of houses on the market that were not purchased, followed by the Klang Valley, Perak, and Penang. Purchasing and owning a dream house is the ambition of all men and women, however, there are a lot of houses and properties which are overpriced and unaffordable by most Malaysians, creating a problem whereby the rate of house ownership is getting lower in Malaysia over the years. In this work, three machine learning algorithms, namely linear regression, random forest, and extreme gradient boosting were employed to predict housing prices. Principle component analysis was also utilised in this work to reduce the dimensionality of the dataset, and cross-validation was also performed to ensure that the models are not overfitted. The aim of this work is to develop a predictive regression model that is most suitable to be used to predict the house price in Malaysia. The dataset used in this work was obtained from Kaggle, which is a dataset presenting the house prices of King County in the U.S. State of Washington. From the results obtained, extreme gradient boosting demonstrated the best performance compared to the other algorithms, with the lowest MAE, RMSE and MAPE. This indicates that the utilisation of the machine learning approach in the prediction of house prices is effective and further investigations should be carried out to obtain more consistent findings.

I. Introduction

The affordability of housing remains to be a persistent issue in Malaysia, particularly for those dwelling in major cities, including the Klang Valley, Penang, and Johor Bahru (Hassan et al., 2021). As reported by the Financial Surveillance Department of the New Straits Times, as of 2019, there were 73 percent of unsold residential properties that were unaffordable by Malaysian citizens, with Johor having the highest record of unsold houses, followed by the Klang Valley, Perak and Penang (New Straits Times, 2019). In addition, a survey conducted by Demographia reveals that the median yearly household income in Malaysia exceeded about 5.5 times the median housing price in 2014 (The Malay Mail, 2014). Affordability of housing is one of the key factors describing the growth and socioeconomic equilibrium of a nation. Moreover, housing affordability seeks to ensure that housing is affordable for all categories of wage earners, including those with low, moderate, and high incomes (Hassan et al., 2021).

A home is a necessity for a family's survival and a fundamental requirement for all humans. Ownership of a house is a major life goal that everyone strives for. The price of a house is the most important factor when deciding to purchase a house. In the interim, purchasers also assess the property's location, accessibility, and amenities, among other factors. The price of a property may increase if adjacent highways and public transportation are improved, and it will continue to grow if the property owner or broker bids up the price in response to market demand (Liew & Haron, 2013).

In contrast to the Light Rail Transit 3 project, which was intended to link the capital city and its suburbs, the Klang Valley’s housing and real estate markets are growing at an unhealthy rate. Undoubtedly, developers compete for sites close to designated stations, claiming that these are desirable locations. In addition, property prices have risen exponentially alongside the inflation rate, crude oil price, and other factors. In actuality, the high value of home prices caused buyers to delay the purchase of a home, be compelled to seek alternatives to their preferences, or endure a large housing loan. During the year of 2009 to 2011, Klang Valley housing prices increased dramatically, rising by 17 percent, while transaction volume increased by 14 percent (The Edge Markets, 2012).

The rate of homeownership in Klang Valley is directly influenced by the price of housing in the region. Therefore, to prevent an increasing number of Malaysians from losing the chance to purchase a reasonably priced home, it is crucial to conduct in-depth research on the issue. A computer aided design (CAD) system can facilitate the prediction of housing prices using machine learning algorithms. This will facilitate real estate agents or house seller to better predict the reasonable price of the property, as well as house buyers, especially the first timers to be able to gauge how much is the price of the type of house that they are looking for. This will ensure that the buyers will not be paying way more extra pennies for what they are getting, and also to lessen their burdens in suffering a huge amount of housing loan.

Machine learning is one of the domains of artificial intelligence (AI) and computer science which allows the machine to learn and improve from experience and predicting outcomes accurately without being explicitly programmed. This paper aims to develop a machine learning model that can predict housing prices with high accuracy in order to increase the rate of homeownership in Klang Valley. A few techniques will be used and compared in this paper, including linear regression (LR), random forest (RF), and extreme gradient boosting (XGBoost) regression.

II. Objectives

To develop a machine learning model that can predict housing prices with high accuracy.

III. Related Works

Park and Bae (2015) developed a housing price prediction model using machine learning models such as C4.5, RIPPER, Naïve Bayes (NB) and AdaBoost algorithms. The dataset used was based on the townhouses in Fairfax County, Virginia, obtained from the Metropolitan Regional Information Systems (MRIS). From the results of the study, it was shown that the RIPPER algorithm demonstrated the best performance in terms of accuracy, as it consistently outperforms the other models in the valuation of housing prices. Banerjee and Dutta (2017) carried out a study to predict whether the house prices in India will rise or fall using machine learning classification techniques. The performance of the models was evaluated based on their accuracy, precision, specificity, and sensitivity. It was shown that random forest (RF) demonstrated the best performance with 86% accuracy, 79% precision, 82% sensitivity and 83% specificity.

Ghosalkar and Dhage (2018) created a linear regression (LR) model to predict house prices in Mumbai, India. The results of the study showed that LR can be used to estimate house prices with a minimum prediction error of 0.3713. Phan (2019) developed a machine learning model to forecast the house prices of Melbourne, Australia. The dataset was taken from Kaggle and the dataset consists of transactional variables, the location details of the house, as well as the features of the house such as the type of the house, the number of bedrooms and bathrooms. Principle component analysis (PCA) and the stepwise method were used for feature selection. The results showed that among LR, polynomial regression (PR), regression tree (RT), neural network (NN) and support vector machine (SVM), stepwise and tuned SVM hybrid model demonstrated the best performance in terms of accuracy.

Vineeth et al. (2018) utilised LR, multiple linear regression (MLR) and NN to predict house prices. The NN demonstrated the best performance as it achieved the lowest root mean squared error (RMSE). Wang and Wu (2018) proposed a RF model to estimate house prices, and LR was used as the benchmark. The housing data used was based on the single-family houses in the county of Arlington, Virginia, USA. The RF model demonstrated a consistently higher R-square value and lower RMSE as compared to LR. The housing prices of the city of Krasnoyarsk, Russia was estimated by Koktashev et al. (2019) using the RF, ridge regression and LR models. According to the mean absolute error (MAE) of the models, it was suggested that RF has the best performance as it has the least MAE value.

Madhuri et al. (2019) used several regression algorithms such as MLR, ridge regression, LASSO regression, elastic net regression, extreme gradient boosting (XGBoost) and Ada boosting (AdaBoost) in the prediction of house prices in USA. The aim of the study is to assist real estate agents or house seller to estimate the selling price of the property accurately, as well as to help citizens to predict the perfect time to buy a house. Based on the results of the study, it can be concluded that XGBoost demonstrated the best performance as it has the highest accuracy score and the lowest mean squared error (MSE) and RMSE as compared to all the other algorithms. Ahtesham et al. (2020) used XGBoost to predict house prices in the city of Karachi, Pakistan and the model achieved an accuracy of 98%.

Jha et al. (2020) studied the performance of XGBoost, CatBoost, RF, LASSO regressor and voting regressor in the prediction of house prices. The dataset used was obtained from the Florida's Volusia Country Property Appraiser website. Based on the R-square value, MSE and MAE of the machine learning models, it was shown that XGBoost demonstrated superior performance as compared to the other models. Thamarai and Malarvizhi (2020) developed a model using decision tree (DT) regression and MLR to predict the housing prices of a small town in West Godovari, Andhra Pradesh, India. The authors suggested that DT has a better performance compared to MLR as it has lower MAE, MSE and RMSE.

Dabreo et al. (2021) employed XGBoost, RF, DT and LR regression models on two different datasets, which are the Boston House Price Dataset and Melbourne Dataset retrieved from Kaggle. Based on the RMSE mean, RMSE standard deviation and mean cross-validation score, XGBoost demonstrated the best performance for both datasets, with the lowest RMSE mean and RMSE standard deviation value and highest mean cross-validation score. Ho et al. (2021) estimated the housing prices in Hong Kong using three machine learning algorithms, namely SVM, RF and XGBoost. The dataset used consists of 40,000 housing transactions over a period of 18 years. XGBoost exhibited the highest performance with the highest R-square value, and lowest MSE, RMSE and mean absolute percentage error (MAPE) value as compared to all the other algorithms.

Imran et al. (2021) explored a wide range of machine learning regression techniques to predict house prices in the city of Islamabad, Pakistan. The results shown that among all the regression models studied, support vector regression (SVR) demonstrated the best performance based on the MAPE, MAE and RMSE values. Adetunji et al. (2021) evaluated the performance of RF in the prediction of housing prices in Boston. The predicted house prices using the model were compared with the actual prices, and it was discovered that the difference between the values were within the accepted error margin, which is ± 5.

A summary of all the previous related works reviewed is shown in Table 1 below. From the table, it can be seen that in most of the studies, XGBoost demonstrated the best performance among all the other models. Besides, it was also shown that linear regression is the most commonly used method by researchers in the prediction of housing prices. Random forest has been shown to be the second most commonly used method, this is because random forest can capture hidden non-linear relationship between the dependent variable and the independent variables, hence demonstrating higher accuracy in regression problems (Wang & Wu, 2018).

Table 1. A summary of the previous related works reviewed.

Author(s)	Regression Technique(s)	Best Technique	Accuracy
Park & Bae (2015)	C4.5, RIPPER, Naïve Bayes, Ada Boosting	RIPPER	
Banerjee & Dutta (2017)	Artificial Neural Network, Support Vector Machine, Random Forest	Support Vector Machine	Accuracy: 82% Precision: 75% Sensitivity: 84% Specificity: 81%
Ghosalkar & Dhage (2018)	Linear Regression	Linear Regression	MSE: 0.3713
Phan (2018)	Decision Tree, Neural Network, Support Vector Machine	Stepwise and tuned Support Vector Machine	Evaluation Ratio: 0.56
Vineeth et al. (2018)	Linear Regression, Multiple Linear Regression, Neural Network	Neural Network	RMSE: 2.1905
Wang & Wu (2018)	Random Forest, Linear Regression	Random Forest	R2: 0.701 RMSE: 352.066
Koktashev et al. (2019)	Random Forest, Ridge Regression, Linear Regression	Random Forest	MAE: 209143
Madhuri et al. (2019)	Multiple Linear Regression, Ridge Regression, LASSO Regression, Elastic Net Regression, Ada Boosting, XGBoost	XGBoost	Accuracy: 0.9177 MSE: 12037006088 RMSE: 10971390390
Ahtesham et al. (2020)	XGBoost	XGBoost	Accuracy: 98% MAE: 22502.082
Jha et al. (2020)	LR, SVR, DT, RF, XGBoost, LASSO, Voting, CatBoost	XGBoost	R2: 0.97 MSE: 8.17 MAE: 4.03
Thamarai & Malarvizhi (2020)	Decision Tree Regression, Multiple Linear Regression	Multiple Linear Regression	MAE: 1.953 MSE: 6.065 RMSE: 2.463

Dabreo et al. (2021)	XGBoost, Random Forest, Decision Tree, Linear Regression	XGBoost	*Boston Dataset:* RMSE Mean: 3.06 RMSE SD: 0.75 Mean CV Score: 0.88 *Melbourne Dataset:* RMSE Mean: 299880.4 RMSE SD: 23019.5 Mean CV Score: 0.779
Ho et al. (2021)	Support Vector Machine, Random Forest, XGBoost	XGBoost	R2: 0.90365 MSE: 0.00793 RMSE: 0.08903 MAPE: 0.32251%
Imran et al. (2021)	LR, BRR, SVR, SGDR, Elastic Net, GBR, LASSO, RF, PAR, Theil-Sen	SVR	MAPE: 1918.4957 MAE: 8595.6057 RMSE: 18209.5558
Adetunji et al. (2022)	Random Forest	Random Forest	Error margin: ± 5

IV. Methods

i. Dataset

The dataset was obtained from Kaggle and it presents the house prices of King County in U.S. State of Washington. The dataset consists of 21 attributes and 21,613 instances. The attributes of the dataset include id, date, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, lat, long, sqft_living15, sqft_lot15, and price as the dependent variable. The details of the attributes are described in Table 2 below.

Table 2. The descriptions of the attributes of the dataset.

Attributes	Description	Data Type
id	The ID of the house.	Nominal
date	The date of the day when the house was sold.	Date
price	The price of the house.	Ratio
bedrooms	The number of bedrooms in the house.	Ratio
bathrooms	The number of bathrooms in the house.	Ratio
sqft_living	The square footage of the house.	Ratio
sqft_lot	The square footage of the lot.	Ratio
floors	The number of levels in the house.	Ratio
waterfront	This describes whether the house has a view to a waterfront.	Nominal
view	The number of times the house has been viewed.	Ratio
condition	The condition of the house.	Ordinal
grade	The overall grade given to the house (according to the grading system of King County).	Ordinal
sqft_above	The square footage of the house apart from the basement.	Ratio
sqft_basement	The square footage of the basement.	Ratio
yr_built	The year that the house was built.	Year
yr_renovated	The year that the house was renovated.	Year
zipcode	The zip code of the location of the house.	Nominal
lat	The latitude coordinate of the house.	Latitude
long	The longitude coordinate of the house.	Longitude
sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors.	Ratio
sqft_lot15	The square footage of the land lots of the nearest 15 neighbors.	Ratio

ii. Machine Learning Algorithms and Techniques

In this paper, three machine learning algorithms, namely Linear Regression (LR), Random Forest (RF) and Gradient Boosting (XGBoost) were used to predict house prices in King County. LR was used because from all the literatures reviewed, this is the most commonly used method in predicting house prices. Besides, LR is also the most commonly used machine learning technique for regression problems. XGBoost was chosen because in most of the previous related works, XGBoost had the best performance in predicting housing prices. Hence, the performance of XGBoost will be verified in this paper. Moreover, XGBoost is said to be one of the most powerful machine learning techniques for both classification and regression problems. RF was chosen because from all the literatures reviewed, RF comes second in performance after XGBoost, and it is said to be able to capture hidden non-linear relationship between the dependent variable and the independent variables, hence demonstrating higher accuracy in regression problems (Wang & Wu, 2018).

Principle Component Analysis (PCA), a feature reduction method will also be performed on the dataset to reduce the dimensionality of the dataset. This is because if the dataset is huge, some of the variables might not have much influence on the dependent variable, hence this will decrease the performance of the prediction model. Besides, having a huge dataset may unnecessarily increase the computational time of the prediction model. Hence, PCA can reduce the dimensionality of the dataset and improve the performance of the prediction model. Furthermore, in this paper, cross validation will also be performed using the machine learning models. Cross validation is very useful in assessing the effectiveness of the predictive model, and also to avoid overfitting of the model.

iii. Performance matrices

The Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) will be used as the performance evaluation matrices to evaluate the performance of the regression models.

iv. Implementation of the Models

The dataset pre-processing steps and machine learning models will be built and implemented using R Studio version 4.1.3 on an ASUS Vivobook with 8 GB RAM.

V. Dataset Preparation

First and foremost, the id and date columns of the dataset were removed since they will not have any impact on the dependent variable. Then, the zipcode, lat, and long columns of the dataset were also removed. These three variables are the variables indicating the exact location of the house. However, they have very similar values and hence, they do not have high impact on the dependent variable. This is further verified by the correlation coefficients of the attributes with the dependent variable (price). The correlation coefficients of zipcode, lat and long with price are -0.05, 0.31 and 0.02, respectively.

Then, it is checked whether there are any missing values in the dataset. The dataset has no missing values but there were a few discrepancies in the dataset. For examples, from the summary of the dataset, it can be seen that there were a few houses with zero bedrooms and zero bathrooms, which do not really make sense. This might be a mistake made by the real estate agent who manually key in the values of the dataset. Hence, records with zero bedrooms and zero bathrooms were removed and there were 21,597 records left in the dataset.

Next, the waterfront variable which has categorical arguments “Yes” and “No” was encoded. After encoding, 1 is “Yes” and 0 is “No”. Besides, yr_built and yr_renovated indicate the year that the house was built and the year that the house was renovated, respectively. Since the year of these variables do not really make sense on impacting the dependent variable, these columns were modified by subtracting the years by 2022. After subtracting, the values obtained will be the number of years since the house was built or renovated. Hence, new columns were created after subtracting, namely yr_since_built and yr_since_renovated, respectively, and the original columns were dropped. However, some of the houses have not been renovated, for those houses which have never been renovated, they were indicated with “0” in the dataset, and hence the subtracted values will be “2022”. Thus, these “2022” will be replaced by the number of years since the house was built, indicating that the house was never renovated since it was built.

After data preparation, the dataset was randomly split into the training set and the test set with the ratio of 0.8 / 0.2, which means 80% of the dataset will be used as the training set and 20% will be used as the test set. After splitting the dataset, feature scaling was applied on both of the training set and the test set. Feature scaling plays an important role in normalising the range of independent variables or features of data to improve the performance of the prediction models and get good accuracy. This is because without feature scaling, the machine might learn that features with larger values are more superior in affecting the outcome of the dependent variable.

VI. Algorithms Model Implementation & Model Validation

Three machine learning regression algorithms, namely LR, RF and XGBoost were used to predict housing prices using the King County House Price Dataset. The dataset was cleaned and pre-processed as mentioned in the previous section. PCA was also applied on the dataset to reduce the dimensionality of the dataset. The regression models were applied on the original dataset and also the PCA-reduced dataset. After that, 10-fold cross validation and hyperparameter tuning were applied to LR and XGBoost using the original dataset and PCA-reduced dataset. Table 3 displays the results of the performance of the algorithms on the original dataset and PCA-reduced dataset before cross validation, while Table 4 illustrates the results of the performance of the algorithms on the original dataset and PCA-reduced dataset after cross validation and hyperparameter tuning.

Table 3. Experimental results of the machine learning models on the original dataset and PCA-reduced dataset.

Machine Learning Models	MAE	RMSE	MAPE
LR	143436.0	207746.2	0.3002
RF	118383.1	178232.5	0.2407
XGBoost	122659.1	187291.9	0.2435
PCA-LR	166340.4	237200.9	0.3501
PCA-RF	164078.6	253325.7	0.3339
PCA-XGBoost	155526.9	255051.5	0.3078

Table 4. Experimental results of the machine learning models on the original dataset and PCA-reduced dataset after cross validation.

Machine Learning Models	MAE	RMSE	MAPE
LR	143436.0	207746.2	0.3002
XGBoost	117145.1	176487.4	0.2297
PCA-LR	166340.4	237200.9	0.3501
PCA-XGBoost	158001.0	244306.6	0.3226

VII. Analysis and Recommendations

Based on Table 3 and Table 4, inference of the accuracy of all the algorithms can be made and the best and most suitable models to be used for house price prediction can be determined.

Table 3 shows the MAE, RMSE and MAPE of the models performed against the original dataset and PCA-reduced dataset without performing cross-validation. For the original dataset, random forest showed the best accuracy with the lowest MAE, RMSE and MAPE, followed by XGBoost. Although the MAE of XGBoost is higher compared to linear regression, the RMSE and MAPE is much lower. Surprisingly, the accuracy of the models against the PCA-reduced dataset is much lower compared to the original dataset. This is because PCA plays a role in reducing the dimensionality of the dataset, hence facilitate better prediction of the models as some of the variables in the dataset have minimum to zero effects on the dependent variable. Hence, by removing those variables should result in better accuracy. However, based on Table 3, as for the PCA-reduced data, XGBoost showed the best performance compared to linear regression and random forest, with the lowest MAE, RMSE and MAPE.

On the other hand, the MAE, RMSE and MAPE of the linear regression and XGBoost models against the original dataset and PCA-reduced dataset after cross-validation and hyperparameter tuning. In general, XGBoost performed better compared to linear regression against both the original dataset and PCA-reduced dataset. The MAE, RMSE and MAPE of the XGBoost is much lower than linear regression. The results are within our expectations since according to most of the literatures reviewed, XGBoost demonstrated the best performance among all the algorithms tested. All of these model predictions in Table 4 were examined for overfitting using cross-validation and hyperparameter tuning to evaluate these values. Therefore, these results are exceedingly accurate.

Similar to the results of the models without performing cross-validation, the accuracy of the models against the original dataset is better compared to the PCA-reduced dataset. This might be due to the PCA is not suitable to be performed on this dataset. Therefore, in the future, researchers who want to use this dataset to predict the house price of King County can consider using the linear discriminant analysis (LDA) and kernel-PCA to reduce the dimensionality of this dataset.

Generally, comparing the results obtained from this work with the previous related works reviewed, it can be said that the accuracy of XGBoost and random forest are better compared to the works of Koktashev et al. (2019), Madhuri et al. (2019) and Dabreo et al. (2021).

However, as mentioned previously, the outcome obtained from this work is consistent with the outcome of the previous related works reviewed. That is, generally, XGBoost has the best performance in house price prediction, followed by random forest. Therefore, it can be said that XGBoost is the most suitable algorithm to be used for house price prediction. This important findings in the AI domain can be recommended to the real estate agents or house seller in order to facilitate them in setting the house prices accurately based on the characteristics, design and location of the house.

VIII. Conclusion

In this paper, several predictive machine learning regression models, namely linear regression, random forest, and XGBoost, were built using the King County house price dataset. The models were tested again two datasets, the original dataset and the PCA-reduced dataset. 10-fold cross-validation was also performed using the models against the original dataset and PCA-reduced dataset. In overall, the results show that XGBoost demonstrated the best performance in predicting house price after cross-validation against the original dataset. The results obtained is consistent with most of the literatures reviewed. This means that XGBoost is effective in predicting house prices and this algorithm can be recommended to property developers, real-estate agents, and house seller.

In this work, it is discussed that PCA might not be suitable to be used against this dataset. Hence, in the future work, LDA and kernel-PCA will be used to reduce the dimensionality of this dataset. Besides, since both XGBoost and random forest are bagging methods and proven to be effective in predicting house prices, AdaBoost, which is another bagging method can also be tested against the dataset.

IX. Bibliography

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2021). House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science, 199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

Ahtesham, M., Bawany, N. Z., & Fatima, K. (2020). House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan. *Proceedings - 2020 21st International Arab Conference on Information Technology, ACIT 2020*. https://doi.org/10.1109/ACIT50332.2020.9300074

Banerjee Debanjan and Dutta Suchibrota. (2017). IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) - 2017 : 21st & 22nd September 2017. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2998–3000.

Dabreo, S., Rodrigues, S., Rodrigues, V., & Shah, P. (2021). Real Estate Price Prediction. *International Journal of Engineering Research & Technology (IJERT), 10*(04). https://doi.org/10.17577/IJERTV10IS040322

Ghosalkar, N. N., & Dhage, S. N. (2018). Real Estate Value Prediction Using Linear Regression. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, 1–5. https://doi.org/10.1109/ICCUBEA.2018.8697639

Hassan, M. M., Ahmad, N., & Hashim, A. H. (2021). A review on housing affordability in malaysia: Are we doing fine? *Malaysian Journal of Consumer and Family Economics, 26*(July), 181–206.

Ho, W. K. O., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research, 38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Imran, I., Zaman, U., Waqar, M., & Zaman, A. (2021). Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data. *Soft Computing and Machine Intelligence Journal*, *1*, 2021. https://doi.org/10.22995/scmi.2021.1.1.03

Jha, S. B., Babiceanu, R. F., Pandey, V., & Jha, R. K. (2020). *Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study*. http://arxiv.org/abs/2006.10092

Koktashev, V., Makeev, V., Shchepin, E., Peresunko, P., & Tynchenko, V. V. (2019). Pricing modeling in the housing market with urban infrastructure effect. *Journal of Physics: Conference Series*, *1353*(1). https://doi.org/10.1088/1742-6596/1353/1/012139

Liew, C., & Haron, N. A. (2013). FACTORS INFLUENCING THE RISE OF HOUSE PRICE IN KLANG VALLEY Related papers A ST UDY ON CUST OMER PREFERENCES AND PERCEPT IONS ON QUALIT Y AND SERVICES OF R… Edit or IJRET AN ANALYSIS OF DRIVERS AFFECT ING T HE IMPLEMENTAT ION OF Edit or IJRET A ST UDY ON. *IJRET: International Journal of Research in Engineering and Technology*, *2*(10), 261–272. http://www.ijret.org

Madhuri, C. H. R., Anuradha, G., & Pujitha, M. V. (2019). House Price Prediction Using Regression Techniques: A Comparative Study. *6th IEEE International Conference on &amp;Amp;Amp;Amp;Amp;Quot;Smart Structures and Systems&amp;Amp;Amp;Amp;Amp;Quot;, ICSSS 2019*, 1–5. https://doi.org/10.1109/ICSSS.2019.8882834

Malaysian homes 'seriously unaffordable', says BNM official. (2019). Retrieved 21 July 2022, from https://www.nst.com.my/business/2019/10/532940/malaysian-homes-seriously-unaffordable-says-bnm-official

Park, B., & Kwon Bae, J. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040

Phan, T. D. (2019). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. *Proceedings - International Conference on Machine Learning and Data Engineering, ICMLDE 2018*, 8–13. https://doi.org/10.1109/iCMLDE.2018.00017

Property prices increase at slower pace. (2012). Retrieved 21 July 2022, from https://www.theedgemarkets.com/article/property-prices-increase-slower-pace

Report shows Malaysian homes more unaffordable than in Singapore, Japan and the US. (2022). Retrieved 21 July 2022, from http://www.themalaymailonline.com/malaysia/article/report-shows-malaysian-homes-more-unaffordable-than-in-singaporejapan-and#OlEfCWe5eXzP0qid.97

Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering and Electronic Business*, *12*(2), 15–20. https://doi.org/10.5815/ijieeb.2020.02.03

Vineeth, N., Ayyappa, M., & Bharathi, B. (2018). House Price Prediction Using Machine Learning Algorithms. In *Communications in Computer and Information Science* (Vol. 837). Springer Singapore. https://doi.org/10.1007/978-981-13-1936-5_45

Wang, C., & Wu, H. (2018). A new machine learning approach to house price estimation. *New Trends in Mathematical Science*, *4*(6), 165–171. https://doi.org/10.20852/ntmsci.2018.327