Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia

Thuraiya Mohd, Suraya Masrom, Noraini Johari

Abstract— This paper demonstrates the utilization of machine learning algorithms in the prediction of housing selling prices on real dataset collected from the Petaling Jaya area, Selangor, Malaysia. To date, literature about research on machine learning prediction of housing selling price in Malaysia is scarce. This paper provides a brief review of the existing machine learning algorithms for the prediction problem and presents the characteristics of the collected datasets with different groups of feature selection. The findings indicate that using irrelevant features from the dataset can decrease the accuracy of the prediction models.

Index Terms: House pricing, machine learning, prediction, real dataset.

I. INTRODUCTION

In the era of Industry 4.0, many urgent issues in industries can be effectively solved with big data techniques including machine learning. Research shows that machine learning tools have been very useful in solving many problems of prediction and classification with wide spectrums of application including medical diagnosis, business analysis, fraud detection, and handwriting recognition. Unfortunately, the utilization of machine learning in real estate applications mainly in the Malaysian context is relatively few and far between.

Evaluating property prices/values is extremely important for real estate, the stock market, the tax sector, the economy, and the size of buyers' and sellers' wallets. While the researchers are reasonably spot-on with current predictions, the current methodologies limited by the scope of data of the current systems in the real estate industry should be taken into consideration. Provided with an estimated price for a given subject property, the machine learning tool is always able to accurately see the potential by considering the significant factors. However, to use the machine learning tool in the real estate industry, the determination of significant factors is crucial; requiring pre-processing and exploration of the collected datasets. The accuracy of the results produced by the machine learning model is highly dependent on the dataset pattern, the parameter tunings, and the feature selections.

This paper reports the findings of analyzing machine learning predictors on real dataset of house pricing in an urban area in Malaysia. The objective of this study is to identify which features from the dataset highly contribute to the prediction accuracy when tested with the selected machine learning algorithms. This study will also verify how the significant level of each factor from the features set affects the performances of machine learning algorithms.

This paper is organized as follows: Section II provides the literature on the trends of housing prices and their predictions as well as the utilization of the machine learning tools. The methodology of research is presented in Section III, while a discussion of the results in Section IV is followed with the concluding remarks in the last section.

II. BACKGROUND OF THE STUDY

A. Housing price prediction

There are a few data analysis modelling techniques being implemented in various property pricing research. These modelling techniques are under econometrics whereby modelling is more concerned with the use of a statistical method, or a mathematical method. An example is autoregressive integrated moving average (ARIMA), artificial intelligence (AI), linear regression, artificial neural network (ANN), fuzzy logic, and hedonic price model [1], [2]. A plethora of researchers have reviewed the applications of these model techniques in the real estate field.

The artificial neural network (ANN) is an AI model technique that is already extensively used in property pricing and being widely implemented in other diverse research areas as well. This modelling technique possess highly promising methods and proves significantly efficient for property pricing appraisal research [3]. The most impressively reviewed of this modelling technique is currently applied in a broad range of science disciplines and business fields such as studies in credit card fraud detection, cursive hand writing recognition, loan approvals, real estate analyses and marketing analyses, telecommunications, sound and vibration controls, automotive, and speech recognition [4]. This research bears semblance to other research [5] reviewing the application of the ANN modelling technique in health and medicine, accounting and finance, engineering, marketing, manufacturing and other application fields. These reviews demonstrate the versatility of this modelling technique that is applicable not only for property price research, but in other industries of science disciplines and business.

Consequently, the hedonic price model (HPM) [6] cited studies conducted by [7] that claim that HPM was already

Revised Version Manuscript Received on September 16, 2019.

Thuraiya Mohd, Faculty of Architecture, Planning and Surveying, Universiti Teknologi MARA, Perak Branch, 32610 Seri Iskandar, Malaysia.

Suraya Masrom, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, 35000 Tapah Campus, Malaysia.

Noraini Johari, Faculty of Architecture, Planning and Surveying, Universiti Teknologi MARA, Perak Branch, 32610 Seri Iskandar, Malaysia.

used in real estate appraisals prior to 1954, between 1954 to 1964, and between 1964 to 1977. In [7] insisted that during their study duration, this modelling technique has been practiced on a few classes of real estate properties, like urban bare lands, rural bare lands, multiple-family residences, and single-family residences. A detailed theoretical review of the implementation evolution of the HPM modelling technique in the real estate aspect was presented by [8], [9] in concert with other research [10] that analyzed articles on the implementation of HPM to measure the variables that determine property values. One of the advantages of the HPM modelling technique includes its ability to estimate the prediction values based on the variable choices besides being particularly applicable to property market research. Simultaneously, this modelling technique is capable of being adapted in case studies having relationships between other market goods and external factors such as structure and construction, property internal features, property external amenities, environmental, neighborhood and location.

It is interesting to note that a lot of previous research using the HPM modelling technique compares other modelling techniques that focus on property market pricing in North America, the West, and Europe in particular. Of these, just a few were conducted in the East, such as in Japan [9], Hong Kong [10], and Taiwan [11].

The fuzzy logic system (FLS) is a modelling technique utilizing a multiple criteria decision-making method or tool. Notable articles on the application of FLS articles were published between 1994 to 2014 [12]; most prominently in 2013. Based on reviews, this technique is mostly implemented in the fields of management and business, engineering, and science and technology. In [12] reports a hybrid system where the FLS modelling technique integrates with the analytic hierarchy process (AHP) to form a combined modelling technique.

Similarly, the AHP modelling technique is another multiple criterion decision-making technique that has been used in the real estate research [13]. This paper states that AHP is also suitable for application in different studies namely conflict resolution, economic and planning, manufacturing, accounting and auditing, politics and environmental marketing, and education. In summary, the FLS modelling technique is widely used by previous researchers on a variety of studies, but it is not widely implemented in the real estate aspect as compared with the HPM model.

B. Machine learning

Machine learning is prevalent in solving many kinds of problem domains [14]-[17]. Due to the higher accuracy performance of machine learning compared to statistical methods, the existence of house price predictor based of machine learning has been very attractive to many professionals involved in property valuation. Researchers have started to use a variation of machine learning algorithms and this paper give a review of five common algorithms namely Linear Regression, Random Forest, Decision Tree, Lasso, and Ridge. Linear Regression [18] machine learning focuses on multiple input variable regression, which uses a linear combination of independent variables to estimate a continuous dependent variable, thus making it relevant to problem prediction. Meanwhile, Random Forest [19] is one of the most popular machine learning algorithms and recognized as the most powerful supervised machine learning. It has the capability to perform both regression and classification tasks. This algorithm was introduced based on the forest and trees as the underlying concept. Similar to Random Forest, Decision Tree is another type of machine learning tool covering both prediction and classification problems [20]. These both algorithms are useful for decision analyses to visually and explicitly represent a set of decision-making possibilities. The idea behind the Random Forest creation is to resolve the over-fitting issues in decision tree algorithms.

Multi co-linearity is a condition of near-linear relationships among the independent variables of prediction [21]. Inaccurate estimation of prediction is the main problem appearing with multi co-linearity conditions; thus the Ridge regression was introduced to resolve the problem [22].

Lasso or Least Absolute Selection and Shrinkage Operator is an alternative to Ridge for regularizing with linear regression [23]. With an objective to find sparse solutions for better interpretation, Lasso replaces both co-efficient vector penalties in the ridge algorithm with a different formulation. To summarize, Table 1 provides a comparison of research regarding the five mentioned machine learning algorithms used in this study. ANN as mentioned in the previous section is a kind of deep learning from the family of machine learning but is more suitable for complex dataset problems.

Table 1: A comparison of machine learning house price predictions (2016-2018)

Paper	Linear Regression	Random Forest	Decision Tree	Ridge	Lasso
[20]		✓	✓		
[24]	✓	✓	✓		✓
[25]		✓			
[26]	✓	✓			
[27]		✓		✓	✓
[28]		✓	✓		
[29]		✓			
[30]				✓	✓
[31]	✓	✓			
[32]	✓	✓			

Based on the reviewed literature, Random Forest is the most popular machine learning. This paper demonstrates the utilization of machine learning algorithms in the prediction of housing selling prices on real dataset collected from the Petaling Jaya area, Selangor, Malaysia. To date, literature about research on machine learning prediction of housing selling price in Malaysia is scarce. This paper provides a brief review of the existing machine learning algorithms for the prediction problem and presents the characteristics of the collected datasets with different groups of feature selection. The findings indicate that using irrelevant features from the dataset can decrease the accuracy of the prediction models.

III. METHODOLOGY & RESULTS

A. Dataset

The dataset is a collection of housing prices in 2016 with some attributes or factor variables. As this paper uses machine learning predictions, these variables are called features. Table 2 shows the set of features to develop the prediction model. This study uses 19 attributes or features as independent variables for predicting house prices.

Table 2: Features in the dataset

Features	Description
Selling price	Dispose price/sqf (RM)
Buying price	Transaction price/sqf (RM)
Floor	Floor
GC	Green certificate
MFA	Main floor area
Bed	Number of bedrooms
Distance	Distance to CBD
BC	Building category
Ownership	Own
CA	Category area
AC	Area classification
Floor	Floor
BC	Building category
CLASS	Building classification
Bed	Number of bedroom
Age	Age of the building
Buy	Buyers
Sell	Seller

B. Features selection

Features selection is an important step of machine learning prediction. In this paper, features selection is divided into four groups. First group used all the independent parameters in the training dataset. It is a combination of variables with very weak, weak and strong relationships on the dependent variable sale price. In this paper, the level of relationship is defined as Strong if the coefficient correlation value is between 0.51 to 1.00 and moderate if the value is between 0.3 to 0.5. Otherwise, weak level is between 0.2 to 0.29 and very weak level is between 0.1 to 0.19. Fig. 1 shows the Python heatmap plot of all variables in the dataset.

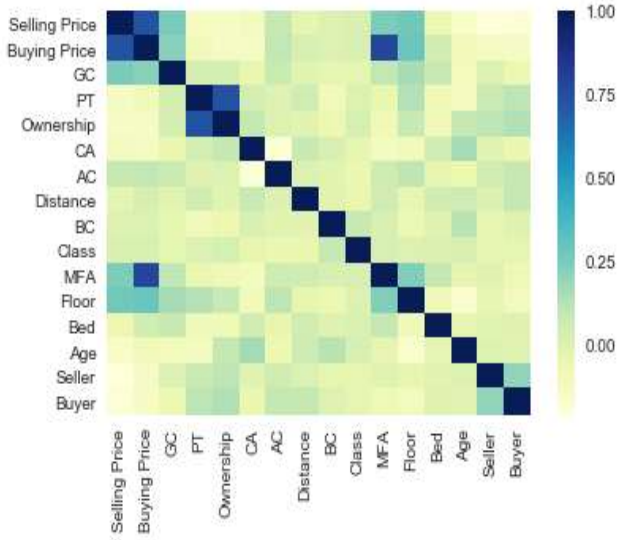

Fig. 1: Correlation level of all features

The following Fig. 2 shows the heatmap plot of weak relationship variables with the selling prices.

Fig. 2: Correlation level of features with values between 0.20-0.29 (weak)

Subsequently, Fig. 3 shows another five variables with very weak relationship with the selling prices. These variables are:

Fig. 3: Correlation level of features with values between 0.1 to 0.19 (very week)

Only the buying price variable is found to have a strong relationship with the selling prices with coefficient value 0.73, as presented in the following Fig. 4.

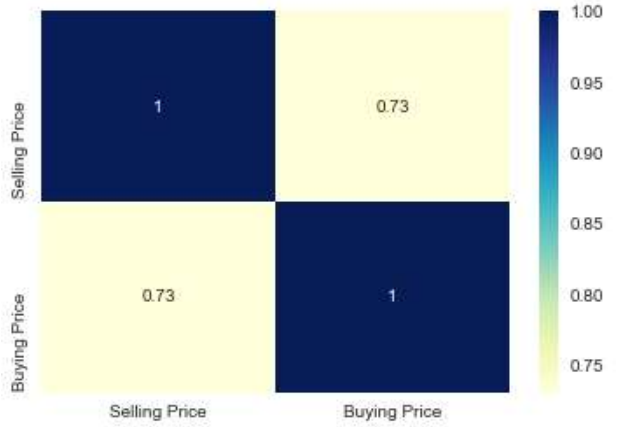

Fig. 4: The correlation level between selling price and buying price

It is interesting to observe the data distribution between the selling price and buying price, thus presented in the following Fig. 5.

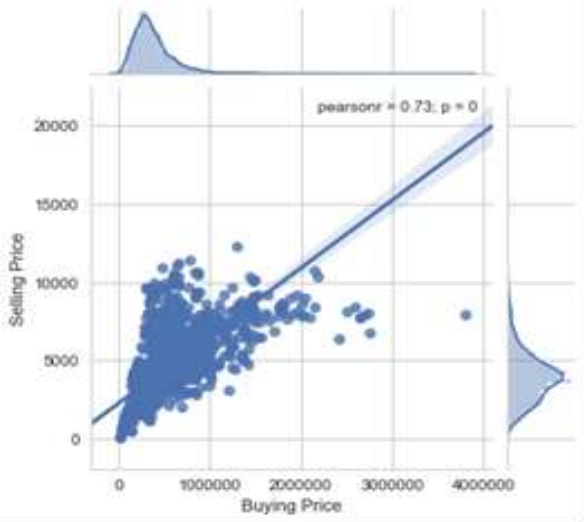

Fig. 5: Data distribution between selling prices and buying prices

There exists a normal distribution trend between selling and distribution prices with very few outliers.

Validation approach

In this study, split training approach is used with a ratio of 80:20 between training and validation respectively.

IV. RESULTS AND DISCUSSION

Table 3 shows the results of R squared and RMSE of five machine learning algorithms on the tested dataset. R squared is a statistical measurement that represents the proportion of the variance for the dependent variable that is explicated by the independent variable/s. The accuracy of each machine learning algorithm can be presented with the R squared.

Table 3: R squared of the five algorithms with different features selection

Algorithm	All	Strong	Weak	Very Weak
Random Forest Regressor	0.991	0.991	0.992	0.817
Decision Tree Regressor	0.986	0.986	0.982	0.804
Ridge	0.683	0.683	0.678	0.428
Linear Regression	0.683	0.683	0.678	0.428
Lasso	0.683	0.683	0.641	0.428

In general, the best accuracy was provided by the Random Forest Regressor followed by the Decision Tree Regressor. A similar result is generated by the Ridge and Linear Regression with a very slight reduction in Lasso. Across all groups of features selections, there is no extreme difference between all regardless of strong or weak groups. It gives a good sign that the buying prices can be solely used for predicting the selling prices without considering other features so as to disseminate model over-fitting. Additionally, a reduction in accuracy is apparent on the very weak features group. The same pattern of results is visible on the Root Square Mean Error (RMSE) for all feature selection groups of all algorithms as shown in the following Table 4.

Table 4: RMSE of the algorithms with different features selection

Algorithm	All	High	Moderate	Weak
Random Forest Regressor	0.044	0.044	0.040	0.202
Decision Tree Regressor	0.056	0.056	0.061	0.209
Ridge	0.267	0.267	0.268	0.358
Linear Regression	0.267	0.267	0.268	0.358
Lasso	0.283	0.283	0.283	0 0.358

RMSE is the standard deviation of prediction errors. It represents the sample standard deviation of the differences between predicted values and observed values (called residuals of independent variables). The lower value is the error signifying the better fit of the algorithms. As presented in Table 1, the inclusion of all features does not affect the fitness of algorithms unless if the model was only utilizing the weak correlation features.

V. CONCLUSION

This paper presents the reviews and findings of using machine learning algorithms for real data of housing prices in the area of Petaling Jaya, Selangor. The researchers demonstrate that feature selection is an important component of machine learning prediction. Two important performances of machine learning prediction are accuracy of prediction, and averaging of errors or fitness, which may be affected according to the feature's selection with different groups of relationship levels. However, these findings are limited to the tested dataset and therefore requires further investigations for different types of problems.

VI. ACKNOWLEDGMENT

The researchers would like to extend gratitude to the Ministry of Education of Malaysia for granting us funding under the Fundamental Research Grant Scheme (FRGS/1/2018/SSO 8/UiTM/02/5) to venture into this research, and for being unwaveringly supportive throughout the term of the research.

REFERENCES

1. V. Limsombunchai, "House price prediction: Hedonic price model vs. artificial neural network," New Zealand Agricultural and Resource Economics Society Conference, 2004, pp. 25–26.
2. N. B. Chaphalkar and S. Sandbhor, "Use of artificial intelligence in real property valuation," Int. J. Eng. Technol., 5(3), 2013, pp. 2334–2337.
3. J. Lee, J. Lee, H. Davari, J. Singh, and V. Pandhare, "Industrial artificial intelligence for industry 4.0-based manufacturing systems," Manuf. Lett., 18, 2018, pp. 20–23.
4. M. Paliwal and U. A. Kumar, "Neural networks and statistical techniques: A review of applications," J. Expert Syst. with Appl., 36(1), 2009, pp. 2–17.
5. R. B. Abidoye and A. P. C. Chan, "Critical review of

hedonic pricing model application in property price appraisal: A case of Nigeria," Int. J. Sustain. Built Environ., 6(1), 2017, pp. 250–259.
6. S. P. Malpezzi, "Hedonic pricing models: A selective and applied review," Hous. Econ. Public Policy, 2001, pp. 67–89.
7. K. W. Chau and K. W. Chau, "A critical review of literature on the hedonic price model," Int. J. Hous. Sci. Its Appl., 74(852), 2003, pp. 3–18.
8. S. Sirmans, D. Macpherson, and E. Zietz, "The composition of Hedonic pricing models," J. Real Estate Lit., 13(1), 2005, pp. 1–44.
9. M. Sasaki and K. Yamamoto, "Hedonic price function for residential area focusing on the reasons for residential preferences in Japanese metropolitan areas," J. Risk Financ. Manag., 11(3), 2018, pp. 2–18.
10. C. K. Wing, S. K. Wong, and L. W. C. Lai, "Hedonic price modelling of environmental attributes: A review of the literature and a Hong Kong case study," Underst. Implement. Sustain. Dev., 2002, pp. 87–110.
11. C. Chen and R. Rothschild, "An application of hedonic pricing analysis to the case of hotel rooms in Taipei," J. Tour. Econ., 16(3), 2010, pp. 685-694.
12. A. Mardani, A. Jusoh, and E. Kazimieras, "Expert systems with applications fuzzy multiple criteria decision making techniques and applications – Two decades review from 1994 to 2014," Expert Syst. Appl., 42(8), 2015, pp. 4126–4148.
13. F. Zahedi, "The analytic hierarchy process-A survey of the method and its applications," Interfaces, 16(4), 1986, pp. 96–108.
14. G. Magesh and P. Swarnalatha, "Attribute reduction and cost optimization using machine learning methods to predict breast cancer," International Journal of Recent Technology and Engineering, 7(6), 2019, pp. 306–308.
15. C. Rajinikanth and S. A. Lincon, "A semi supervised based Hyper Spectral Image (HSI) classification using machine learning approach," International Journal of Recent Technology and Engineering, 7(5S2), 2019, pp. 13–16.
16. S. Singh, M. Kaushik, A. Gupta, and A. K. Malviya, "Weather forecasting using machine learning techniques," SSRN Electron. J., 6, 2019, pp. 38–41.
17. S. Kavipriya and T. Deepa, "Dual Edge Classifier Based Support Vector Machine (DESVM) classifier for clinical dataset," International Journal of Recent Technology and Engineering, 7(6), 2019, pp. 331–338.
18. C. R. Rao and H. Toutenburg, "Linear models," in Linear Models, New York: Springer, 1995, pp. 3–18.
19. A. Liaw, M. Wiener, "Classification and regression by randomForest," R News, 2(3), 2002, pp. 18–22.
20. S. Borde, A. Rane, G. Shende, and S. Shetty, "Real estate investment advising using machine learning," Int. Res. J. Eng. Technol., 4(3), 2017, pp. 1821–1825.
21. D. Disatnik and L. Sivan, "The multicollinearity illusion in moderated regression analysis," Mark. Lett., 27(2), 2016, pp. 403–408.
22. H. Duzan and N. S. B. M. Shariff, "Ridge regression for solving the multicollinearity problem: Review of methods and models," J. Appl. Sci., 15(3), 2015, pp. 392-404.
23. A. B. Owen, "A robust hybrid of lasso and ridge regression," Contemp. Math., 443(7), 2007, pp. 59–72.
24. N. Shinde and K. Gawande, "Valuation of house prices using predictive techniques," Int. J. Adv. Electron. Comput. Sci., 5, 2018, pp. 2393–2835.
25. A. S. Ravikumar, Real estate price prediction using machine learning. Master thesis, Dublin: National College of Ireland, 2017.
26. V. H. Masias, M. A. Valle, F. Crespo, R. Crespo, A. Vargas, and S. Laengle, "Property valuation using machine learning algorithms: A study in a metropolitan-area of Chile," Selection at the AMSE Conferences, 2016, pp. 97-105.
27. S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A hybrid regression technique for house prices prediction," IEEE International Conference on Industrial Engineering and Engineering Management, 2017, pp. 319–323.
28. B. Yang and B. Cao, "Research on ensemble learning-based housing price prediction model," Big Geospatial Data and Data Science, 1, 2018, pp. 1–8.
29. T. Dimopoulos, H. Tyralis, N. Bakas, and D. Hadjimitsis, "Accuracy measurement of random forests and linear regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus," Adv. Geosci., 45, 2018, pp. 377–382.
30. Y. Ma, Z. Zhang, A. Ihler, and B. Pan, "Estimating warehouse rental price using machine learning techniques," Int. J. Comput. Commun. Control, 13(2), 2018, pp. 235–250.
31. K. K. Ganguly, N. Nahar, and B. M. M. Hossain, "A machine learning-based prediction and analysis of flood affected households: A case study of floods in Bangladesh," Int. J. Disaster Risk Reduct., 34, 2019, pp. 283–294.
32. T. Fu, "Forecasting second-hand housing price using artificial intelligence and machine learning techniques," 8th International Conference on Mechatronics, Computer and Education Informationization, 2018, pp. 269-273.