Comparing Econometric Analyses with Machine Learning Approaches: A Study on Singapore Private Property Market

Tingbin Bian, Jin Chen, Qu Feng*, Jingyi Li

Nanyang Technological University, Singapore

Abstract:

We aim to compare econometric analyses with machine learning approaches in the context of Singapore private property market using transaction data covering the period of 1995-2018. A hedonic model is employed to quantify the premiums of important attributes and amenities, with a focus on the premium of distance to nearest Mass Rapid Transit (MRT) stations. In the meantime, an investigation using machine learning algorithms under 3 categories – LASSO, Random Forest and Artificial Neural Networks is conducted in the same context with deeper insights on importance of determinants of property prices. The results suggest that the MRT distance premium is significant and moving 100 meters closer from the mean distance point to the nearest MRT station would increase the overall transacted price by about 15,000 Singapore dollars (SGD). Machine learning approaches generally achieve higher prediction accuracy, and heterogeneous property age premium is suggested by LASSO. Using Random Forest algorithm, we find that property prices are mostly affected by key macroeconomic factors, such as the time of sale, as well as the size and floor level of property. Finally, an appraisal on different approaches is provided for researchers to utilize additional data sources and data-driven approaches to exploit potential causal effects in economic studies.

Keywords: Singapore Property Price; Hedonic Model; Machine Learning Algorithms, Random Forest, Artificial Neural Networks

JEL Classification: C23, C45, R31, R41

---

*Corresponding author. qfeng@ntu.edu.sg, Economics Division, School of Social Sciences, Nanyang Technological University, Singapore. SHHK-04-48, 48 Nanyang Avenue, Singapore 639818. We would like to thank Yang Chen, Yiquan Gu, audiences at *Singapore Economic Review* Conference 2019 and the 7th seminar on Asia and Pacific Economies 2019 at Xi'an Jiaotong-Liverpool University, Suzhou, China. Financial support from the MOE AcRF Tier 1 Grant M4012113 at Nanyang Technological University is gratefully acknowledged.

1 INTRODUCTION

The property market in Singapore is unique as it follows a dual-market structure: a public sector developed by Housing Development Board (HDB) where heavy restrictions and regulations are imposed by the government and a private sector where the prices are fully determined by market forces. 73% of the residential dwelling units in Singapore are government regulated HDB flats which house 80% of all Singaporean citizens (*The Economist*, 2017). Prices of new HDB apartments are set by the government of Singapore and are usually sold to Singaporean citizens at a discounted price. Prices of resale HDB apartments, while not directly controlled, are heavily influenced by government policies that often adjusted depending on the macroeconomic factors affecting the nation (Chan et al, 2012). Despite the small share of the private property market, we focus on the private property market in our paper. The private housing market is rather heterogeneous and composing of transactions by high income households, expatriates and foreign investors. As such, the average household income in the private housing sector is 2.6 times as high as that in the public housing sector (Singapore Department of Statistics, 2015).

It has been well established in the literature that property prices are heavily influenced by the availability of public transportation and connectivity. Property unit price could change with its distance to amenities such as MRT stations and bus stops, however, empirical studies on the magnitude and significance of this impact in the Singapore context are still scant.

This paper aims to compare econometric analyses with machine learning approaches in the context of Singapore's private property market. The data set used in this paper includes all transactions of Singapore's private properties during 1995-2018, with a total number of 516,962 records. With such a relatively large data set, we explore how machine learning approaches could provide additional information that cannot be done by traditional econometric analysis.

We first start with traditional econometric methods, as in the study of Diao et al. (2017) who evaluate the impact of a new MRT line on property prices. With transaction data and an additional collection of distances to amenities and top primary schools across 1995-2018, we quantify the MRT distance premium (MDP). Distribution effects such as whether the properties locate in the Central region and whether the transactions took place before the global financial crisis, and spatial heterogeneity of MDP for households living within different radius ranges from MRT stations are also examined.

Then we move to machine learning approaches to explore whether these new approaches could provide additional insights. The prevalence of big data and the advancement in the area of machine learning have provided economists with diverse and interesting new applications of these technologies. Existing researches of machine learning in the real estate market mainly have been focused on prediction accuracy, but no greater insights on the models and their applications in real estate studies were emphasized, not to mention the MDP in the context of Singapore's private property market. We aim to understand how machine learning tools can be embedded in economic decision-making and what to consider when incorporating these tools in the applied setting. In this paper, several machine learning models are used to extract more insights from our data. Model accuracy and direct impact of distance to the nearest MRT stations on property prices are evaluated and the performance of these data-driven methods on feature selection process is appraised.

In this specific empirical context, using a hedonic specification we find that moving the property 100 meters closer from the mean distance point to the nearest MRT station, one would expect an increase of about 15,000 SGD in the overall transacted price. Using machine learning approaches, we also find that property prices are mostly affected by key macroeconomic factors, such as the time of sale, as well as the size and height of the property. Other important factors include the ease of access to public transportation, living amenities around the property and the age of the property, where the distance to nearest MRT station is the dominant factor among them.

This paper is organized as follows. Section 2 reviews existing literature explicitly on the effects of housing characteristics such as public transport networks on property prices and machine learning applications in the real estate markets. Section 3 introduces data collection and context. Section 4 introduces econometric specifications based on a hedonic model and reports empirical results on premiums of amenities, especially, the MDP. Machine learning algorithms and results are discussed in Section 5, where a comparison between econometric analyses and machine learning approaches is also provided. Section 6 concludes the paper.

2. LITERATURE REVIEW

We first review the empirical studies of the effects of public transport networks on property prices. Investigation into the impact of transport facilities on the property prices has been conducted since 1990s. One of the earliest studies investigating the effects of a new subway line on the housing prices of Toronto finds clear evidence that direct savings in commuting costs have been capitalized into housing values (Bajic, 1993). More recent studies include those in New Jersey of the US (Kim, and Lahr, 2014), Chicago (McMillen and McDonald, 2004), Houston (Pan et al., 2014), Netherlands of Europe (Debrezion et al., 2011), Ireland (Mayor et al., 2008), and China (Zhang et al., 2016), and similar results have been obtained.

However, there are a limited number of studies on the impact of public transport networks on property prices in Singapore. Diao et al. (2017) use a difference-in-differences identification strategy and non-landed private property transaction data and find a 1.6% increase in the average housing price after the opening of Circle

Line. Fesselmeyer and Liu (2018) estimate the impact of an expansion of North East Line on the overall HDB property values and conclude that the expansion increased the overall HDB property values by at least 455 million SGD, 9% of the total construction cost of the expansion.

Secondly, we access the studies which apply machine learning approaches to economics analysis. Machine learning is the method to cope with big data which emphasize the importance of the size and the varieties of data, including touchy data such as images and audio records. Jean et al. (2016) apply neural network to predict the poverty of five African countries according to satellite imagery. Kang et al. (2013) predict hygiene inspection of restaurants according to online review. Machine learning approaches are also widely used for predicting financial crises. For example, Tanaka, Kinkyo and Hamori (2016) use random forest to build an early warning system to signal a county's vulnerability to financial crises, while a multinomial logit model is used instead by Bussiere and Fratzscher (2006). Additional applications include bankruptcy (Tanaka, et al., 2019, Beutel, et al, 2019).

Thirdly, we examine the literature on the application of machine learning techniques to property price prediction. Most of the existing papers has been focused on the prediction accuracy of various algorithms. For example, performances of regression tree, LASSO, random forest, and ensemble algorithms are compared using 10,000 randomly selected owner-occupied dwellings from the 2011 metropolitan sample of the American Housing Survey and the regression tree is found to be the best performing model in terms of prediction accuracy (Mullainathan, et al. 2017). In Singapore's context, ridge regression, LASSO, gradient boosting are compared in terms of their prediction performance using Singapore housing price, and an ensemble algorithm of LASSO and gradient boosting gives the best performance (Lu, et al. 2011). Wang and Zhao (2017) apply Deep Neural Networks to predict HDB housing prices and find that ANN algorithm gives a highest R^2. Li et al. (2018) evaluate the performances of convoluted neural networks (CNN) and long short-term memory (LTSM) on housing price prediction.

Improvement in data collection and expansion of data sources are no doubt highly useful for researchers to predict property price. Creative ways of utilizing new data sources have been also seen for housing price prediction. Bency et al. (2017) propose extracting deep features (DF) from satellite images to capture location and neighborhood effects for London housing price prediction. Places of interest (POI) data from the Google Places web service served as an ancillary data source. As a result, a combination of DF and housing attributes significantly improves prediction accuracy. Besides, Sobolevsky et al. (2017) use a set of anonymized bank card transactions during 2011 in Spain to predict regional economic indices and found them useful to predict major official statistical quantities such as GDP and housing prices. Social media data are utilized to monitor housing rental prices in Shenzhen, China and machine learning algorithms are integrated with the hedonic model to identify determinants and spatial patterns (Hu et al., 2019).

3. DATA AND CONTEXT

The dataset used for this study come from different sources: Real Estate Information System (REALIS), OneMap SG and The Ministry of Education of Singapore (MOE) website. REALIS is an online governmental system maintained by The Urban Redevelopment Authority (URA) that records all private property transactions in Singapore with caveats lodged for residential, commercial and industrial purposes. The data collected from REALIS are limited to residential transactions from 1995Q1 to 2018Q4 and an initial of 516,962 transaction records are collected. Information includes project name, address, number of units, size of the property, type of area, unit price, transacted price, property type, completion year, type of sale, purchaser address indicator, postal code, planning region and area, tenure, transaction date. Appendix 1 presents the detailed definitions of fundamental housing attributes and neighborhood attributes.

OneMap SG is a national map of Singapore developed and maintained by the Singapore Land Authority (SLA) with data provided from other governmental agencies. To get additional information on distances of property to important amenities such as MRT stations, bus stations, food courts, clinics and top primary schools, a list of 3,406 unique project names is generated using R and searched through OneMap SG. The distances from the property to these amenities are then recorded manually. We have also recorded the instances of top Primary Schools within 1km of the property and between 1km to 2km of the property.
We have included 21 primary schools in our list of top primary schools of Singapore (see Appendix 2). The schools are chosen based on the number of MOE education awards received in the past 5 years. MOE does not release official rankings of primary schools. This is done to promote the idea that all public in schools in Singapore are good schools and should be treated equally.

The dataset used for this study reduces to 466,617 observations due to typographical errors and missing information. Simple calculations and rearrangements are made to compute age of the property defined as transacted year – completion year, unit level, tenure level and remaining lease of the property defined as tenure – (transacted year – start date of tenure).

Unit levels are extracted from the addresses as most property addresses are recorded in the uniform format with level information specified right after '#' in the address. For terrace houses, detached houses and semi-detached houses, the levels of these properties are set to 1. Tenure levels are categorized into three categories: 99-year leasehold, 999-year leasehold, and freehold. More than 99% tenures fall into these three categories. The remaining tenures ranging from 100 to 499 years are categorized into 99-year leasehold and tenures ranging from 500 to 998 years are categorized into 999-year leasehold.

Figure 1 presents the number of transactions per year since 1995 in bars. The dotted line represents the moving average of transactions per 2 periods. We note that the sales peaked in 2007, 2010 and 2012 with more than 35,000 transactions in each of these 3 years.

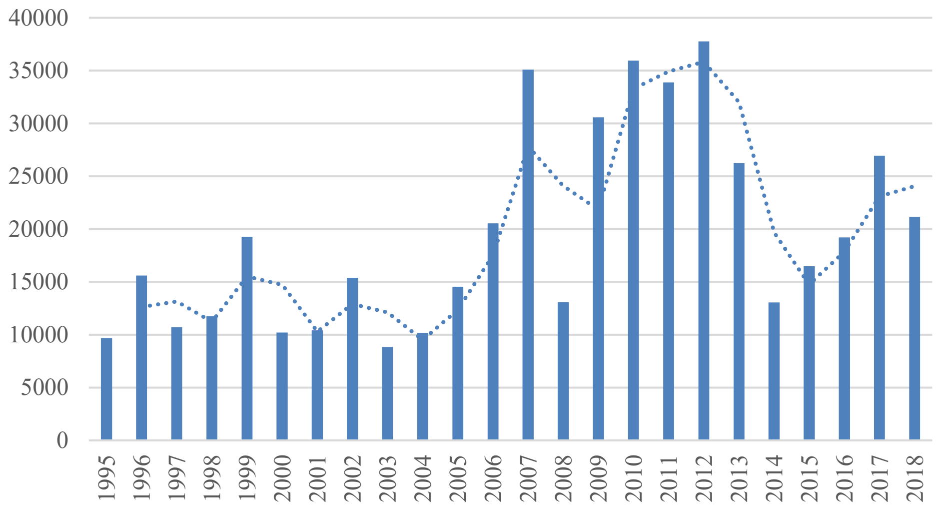

Figure 1 Number of Transactions per Year during 1995-2018

Figure 2 shows the distribution of transactions by property type. Among the whole dataset by property types, condominiums account for 57% of the transactions. The second largest type is apartments, amounting to 25% of all transactions. The third one is executive condominiums, accounting for 9%. The remaining 9% of transactions are split among terrace houses, detached houses and semi-detached houses.

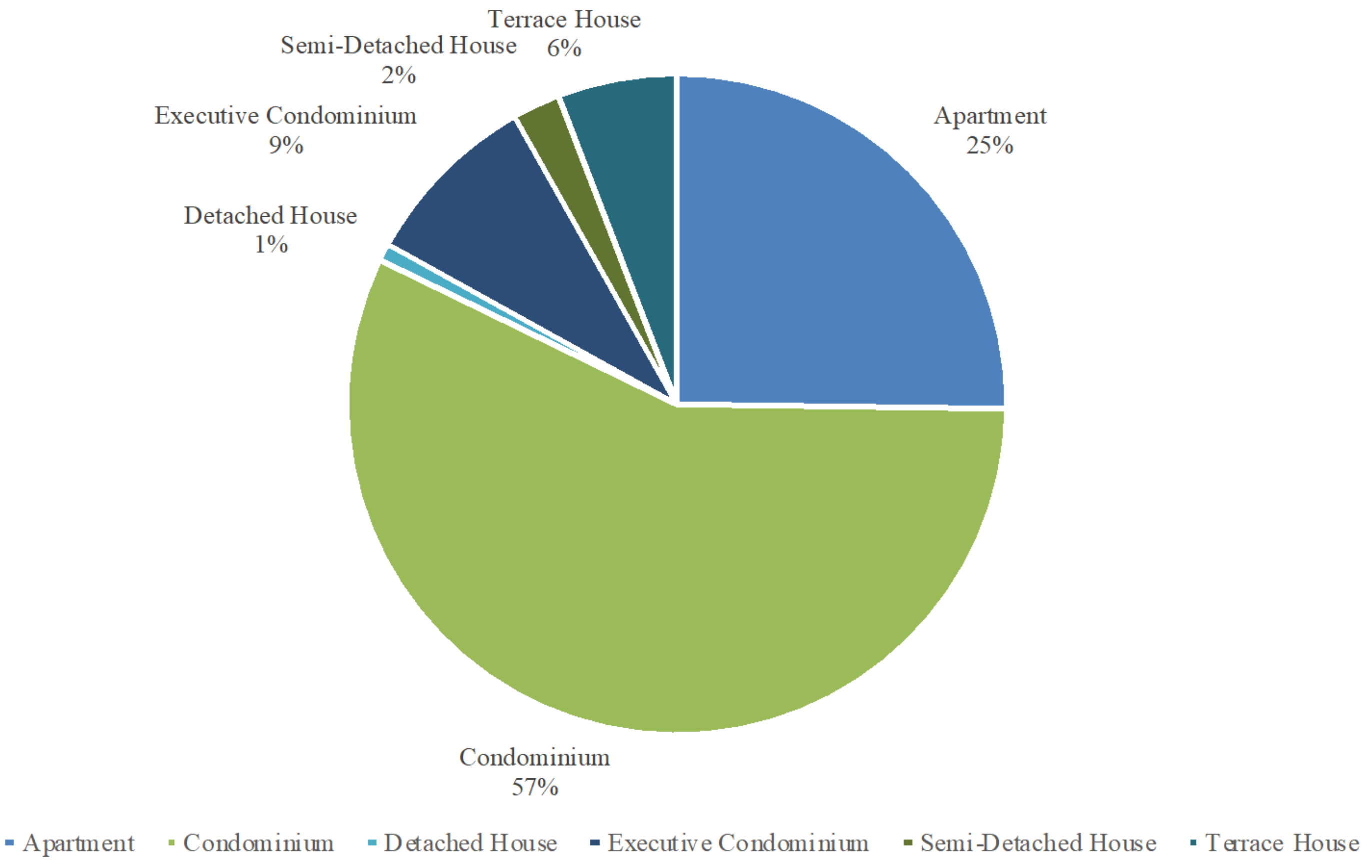

Figure 2 Distribution of Transactions by Property Type

Table 1 shows the number of transactions associated with ranges of distance to nearest MRT stations. Among all transactions, a total of 174 unique MRT stations are recorded and with 87.3% of the transaction locations falling within 1,000 meters to MRT stations. Besides, 16.0% transactions considered in this study are located within 1,000 meters of at least one top primary school and 32.4% transacted properties have at least one top primary school within the distance of between 1,000 and 2,000 meters.

4. ECONOMETRIC MODEL AND MACHINE LEARNING APPROACHES

4.1 ECONOMETRIC MODEL

In this section, we use a hedonic model to quantify the impact of important amenities and distances to MRT stations on the unit price of private properties. The specification used is:

$$y_{it} = \alpha \cdot x1_i + \beta \cdot x2_{it} + \gamma \cdot x3_t + \varepsilon_{it}, i = 1, \dots, N, \tag{1}$$

where y_{it} is the property price per square meter in log form for property i at period t (month), and $x1_i$ denotes a vector of characteristics and amenities of property that are time-invariant, including floor area, floor level, property type dummies whether there is a good primary school/ clinic nearby, distance to nearest MRT stations, etc. Similarly, $x2_{it}$ represent a vector of characteristics that are time varying, e.g., the property age, transaction month, etc. $x3_t$ includes macroeconomic factors that are same for all properties but change over time, e.g., time dummies. For a summary of variables, please refer to Table 2.

The hedonic model is based on Lancaster's (1966) consumer demand theory, which states that consumer utility is derived from the characteristics of goods. Rosen (1974) first introduced the theory of hedonic pricing. Commonly used in housing market studies, the hedonic pricing model links the price of a property to its characteristics, surrounding factors and macro trend. A linear form of hedonic pricing model is adopted in equation (1).

It is worth noting that this data set is a pooled cross-section, instead of a panel data set. Property unit specific factors cannot be controlled in the regression. To identify the impact of amenities on property prices, we assume that unobservable quality variables absorbed in the error term ε_{it} are not correlated with regressors. For instance, more shopping malls and offices are built near the MRT stations and they could simultaneously affect the unit property prices. These variables are not controlled in this study. We use subsamples or location dummies to mitigate the impacts of these factors. Moreover, for robustness we also experiment different specifications, e.g., higher-order terms and more interaction terms.

4.2 MACHINE LEARNING APPROACHES

In this subsection, we briefly explain three popular machine learning algorithms: LASSO, random forest and artificial neural networks. The benchmark model is the regression equation (1), referred to as the OLS (Model 1).

I. LASSO

LASSO applies L_1-norm regularization to penalize parameters of the model in order to avoid overfitting (Tibshirani, 1996), and aims to minimize the loss function specified below:

$$\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|, \qquad (2)$$

where λ is a tuning parameter. The larger λ is, the more parameters will shrink to 0. In equation (2) the intercept is excluded. Besides, Lasso has been proved to be effective to handle nonlinear targets. Therefore, we apply it to explore higher-order and interaction terms related to variable of MRT distance. Equation (2) is referred to as Model (2). In Model (2) same set of variables are used as in the hedonic model (1) and regression in column (1) of Table 3.

To look at how LASSO selects variables, squares terms, cubes terms and interactions of all numerical variables are added to equation (2). This case is referred to as Model (2a). A grid search with 5-fold cross-validation is employed to choose the value of tuning parameter λ, while all other hyperparameters are set to be fixed default values.

II. Random Forest

Random Forest (RF) is a machine learning approach based on decision trees (Breiman, 2001). A decision tree follows a recursive partitioning subject to certain rules, such as to minimize the sum of squared residuals (SSR). Each splitting procedure splits the subset at variable *j* and value *s* into 2 predictor spaces, R_1, R_2. Therefore, the variable *j* and point *s* could be determined in terms of minimizing the SSR on the original subset:

$$\min_{j,s}[\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i \in R_2(j,s)}(y_i - c_2)^2]$$

c_1, c_2 are constant that minimize the SSR in each subset. The splitting process could be repeated on each space until reaching minimum node size, and pruning is required thereafter.

However, decision trees are not as robust as other approaches as a small change in training data can cause a huge change in the model trained. RF employs the bagging and decorrelation of trees to improve the statistical performance of model at a cost of interpretability. Bootstrap aggregation (or bagging) draws *B* datasets each of the same size with replacement from training data, then averages the predictions over bootstrap samples in order to reducing the variance.

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x), \qquad (3)$$

where $\hat{f}^{*b}(x)$ is the prediction over bootstrap sample *b*. Meanwhile, RF reduces the correlation between trees by random selection of the input variables before each split. Typically, $m = q / 3$ of input variables are

randomly selected, where q is the total amount of features. In this way, the correlation among trees is controlled at the minimum.
Gini importance are calculated by averaging the decrease in SSR due to splits over a specific variable over all trees in the model. Meanwhile, random forests also provide a different importance measure, which relies on out-of-bag (OOB) samples. A larger importance value indicates a more important variable (Louppe, 2014).

Two RF models are configured as an attempt to explore the effects of different ways of categorical (or dummy) variable encoding. In the hedonic model (1), we introduced the time effect into through 287 month dummies. However, too many dummy variables may cause the RF approach to lose effectiveness (Au, 2018). In equation (3), one-hot encoding is used for all the categorical variables. It is referred to as Model (3). When label encoding is used in equation (3), it is referred to as Model (3a). One-hot encoding treats all variables as dummies, while label encoding reduces number of variables by assigning random numerical values to each of these variables. For example, for the variable $SALE_i$, values 0, 1 and 2 are assigned to different types of sale. Label encoding does not make sense in hedonic models but works well in an RF model. It does not affect the interpretability of the result as we can interpret $SALE_i > 1$ to a subset of the types of sale. After the treatment, we used a total of 523 features (regressors) for Model (3) and 16 features for Model (3a).

Hyperparameter tuning for RF requires huge computational ability, thus we are only able to focus on the numbers of trees (N). All other parameters are set to be the fixed value, and *max_features* is to be 1/3 of total amount of features. The larger the number N, the more stable the prediction results of the model. After tuning, N for Model (3) is set to be 300 and N for Model (3a) is set to be 200. Please refer to Table 6 for details.

III. Artificial Neural Networks

The advantage of Artificial Neural Networks (ANN) algorithm is that it is possible to approximate any mathematical function given enough data (Ferrari & Stengel, 2015). ANN approach, denoted by Model (4), is applied as a flexible form of regression in our student. Similar in RF, the tuning process in ANN is highly complex. We adopt Keras sequential, a fully connected neural network, in this section. Hyperparameters such as the number of hidden units, hidden layers, batch size, epochs, and dropout rate are tuned, and the ReLU function is selected as the activation function. Our final algorithm has 2 hidden layers with 128 neurons in the first layer and 64 neurons in the second layer. See Figure 3 below for a direct model specification using a graph.

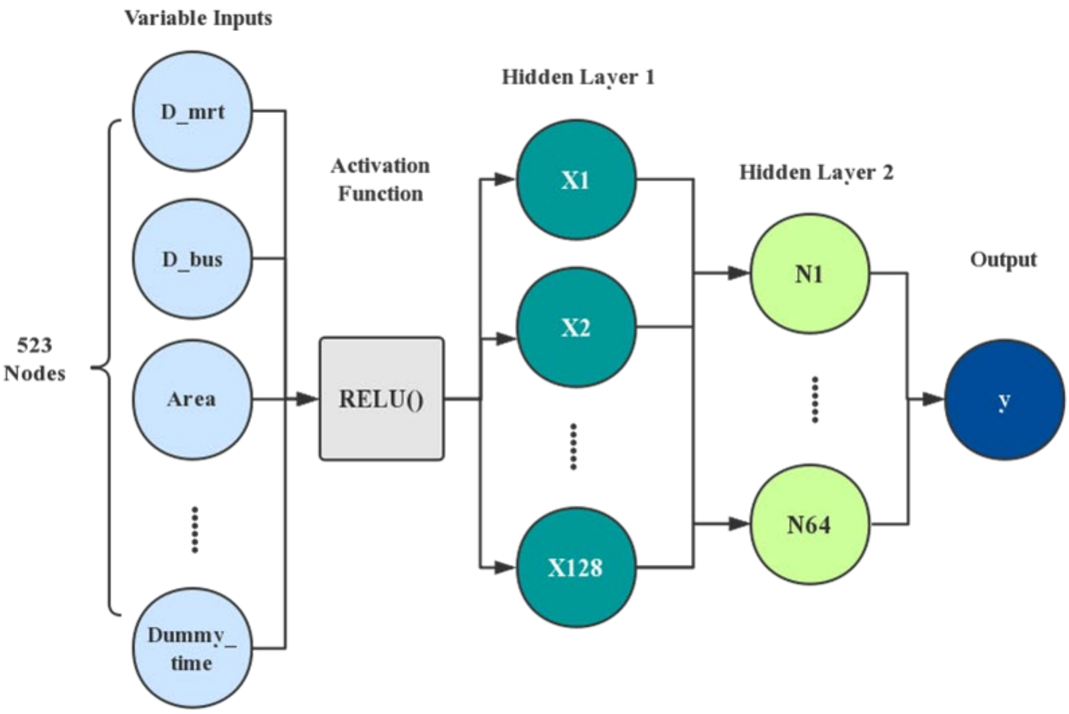

Figure 3 Visualized structure of ANN

5. Empirical Results

5.1 Empirical results of Econometric Analysis

Table 3 reports the effects of amenities on property prices. Since there have been a lot of studies on effects of floor area, tenure, property type on housing prices in the literature, here we focus on the price premiums of important amenities in Singapore context, i.e., distance the nearest MRT (d_{MRT}), distance to nearest bus stop (d_{bus}), distance to the nearest food court (d_{food}), distance to the nearest clinic (d_{clinic}), whether there is a top primary school within 1 km (SCHOOL1) or between 1-2km (SCHOOL2). Other important factors, e.g., type of area (landed vs strata), sale type (new sale, resale and subsale), type of properties (apartment,

detached house, etc), Tenure level (99-years leasehold, 999-years leasehold, freehold), Planning area dummies (38 districts), month dummies of the transactions are also controlled in the regression.
Column (1) of Table 3 reports the estimated coefficient of d_{MRT}, or the elasticity of MRT, using the whole sample. It is -0.067 and is statistically significant. This means that 1% decrease in distance to the nearest MRT station increases the unit price of property by 0.067% on average. In other words, Moving the property 100 meters closer from the mean distance point (603.6) to the nearest MRT station, we expect that the overall transacted price increase by 100/603.6×0.067% =11.1%. Using the information on average values of property price per square meter (10, 441 SGD) and total area (130.0) in Table 2, the 11.1% increase in overall transacted price implies a value of 15,066 SGD.

In column (2), the term of squared d_{MRT} can be used to capture the nonlinear impact of distance to the nearest MRT station on property price. The coefficient of d_{MRT} is -0.026 and that of squared d_{MRT} is -0.004, both statistically significant. The total effect of MRT distance on property price is -0.071 at the mean value of d_{MRT}, which is very close to -0.067 reported in column (1) in the linear case.

In column (3) of Table 3, the interaction term $age \cdot d_{MRT}$ is added to allow effect of MRT distance on property price to change with property age. Now the coefficient of d_{MRT} becomes -0.077. And that of $age \cdot d_{MRT}$ is 0.003, indicating that the impact of MRT distance on property price decays over time by 0.3% per year on average. The overall impact of the distance to the nearest MRT station is -0.065 at the mean value of age.

To capture heterogenous effects over time, we split the whole sample into two subsamples by period: 1995-2008 and 2009-2018. At the end of the 2009 Financial Crisis, there was a spike in property prices in Singapore. To reduce adverse impact caused by the fast growth in property prices, the Singapore government has undertaken significant steps to intervene in the property market (Deng et al, 2018). 10 rounds of cooling measures including amending the mortgage loan-to-value ratios, stamp duties and so on have been adopted since after.
Columns (4) and (5) of Table 3 report the subsample results. It shows that the elasticity of MRT increases in magnitude from -0.056 for the pre-crisis period to -0.071 for the post-crisis period in terms of magnitude. After the financial crisis, the government has imposed strict limitations on mortgage loans to minimize speculative investments to cool surging prices. This could result in a more rational real estate investments after 2008, and buyers value more practical benefits such as closeness to amenities.

In Table 3, Column (6) and (7) also present subsample results for central region and non-central region. The elasticity of MRT becomes -0.090 for properties in central region and -0.039 for non-central region. Different magnitude of elasticity reflects geographic difference in values of land and people's response to this difference. Due to high land costs and importance of the central region of Singapore, the infrastructure is significantly different from that in the rest of Singapore. There is a higher concentration of MRT stations, widespread amenities and no public schools. Thus, property price is more sensitive to the MRT distance in the central region.

We also consider the impact of the type of property by splitting the whole sample into landed and non-landed properties. From our dataset, we found that the mean transaction price of landed properties is 2.12 million SGD which is 76.7% higher than that of non-landed properties. In columns (8) and (9) of Table 3, we observe that the elasticity of MRT for landed properties is -0.010, much smaller in magnitude than -0.067, the elasticity for non-landed properties. Landed property owners are generally more affluent, and they can afford alternative means of private transportation. Hence, MRT distance is relatively less priced in the landed properties.

Table 3 also reports the effects of other amenities and factors, including distance to nearest bus station, food court, clinic and top primary schools. The elasticity of bus is 0.007 is columns (1)-(2) and 0.008 in column (3). Compared with the elasticity of MRT, it has a much smaller magnitude in columns (1)-(9), implying that after controlling for the distance to nearest MRT station, the distance to nearest bus stop has little impact on property price. In addition, in Singapore's private property context, its sign is positive except in column (7), suggesting that the effect of convenience of being close to a bus stop on property price is offset by its effect of congestion and noise of being close to a bus stop. The effects of distance to nearest food court and clinic are very similar to that of distance to nearest bus stop. After MRT is taken into account, being close to a food court or clinic is not necessarily valuable to a private property. The effect of having a top primary school nearby on property price is rather heterogeneous. In columns (1)-(3) using full sample a private property price would increase by only 0.6% if there is a top primary school with 1km. In subsample estimates, its effect could be 2.6% during 1995-2008 in column (4), 1.8% for properties in non-central region in column (7) and 1.9% for landed property in column (8).

An additional robustness check is reported in Table 4. We drop the top and bottom 5% records sorting on the unit prices. The estimates of elasticity of MRT are similar to those in Table 3.

5.2 Empirical Results of Machine Learning Approaches

Prediction Accuracy

Before training the model, we randomly split the dataset into a training set and a test set with a ratio of 4:1. The performances of different machine learning models are presented in Table 5. Both RF models and the ANN models obtain R^2 of 90.4% and 91.2% in test data set in columns (3) and (4), respectively, larger than the 86.1% obtained in the OLS (1). In addition, the RF model (3a) using label encoding on categorical variables has the best performance out of all five models with an out-of-sample R^2 of 97.6%. However, due to its data-driven and black-box characteristics, it is difficult to justify the reasons for the prediction accuracy improvement. [1]

Looking at the accuracy of predictions, we note that LASSO Model (2) is not more accurate than the OLS model (1). As pointed out by Melkumova and Shatskikh (2017), if the relationship between the variables is highly linear and sample size is much greater than the number of features, then the OLS model is likely to give better results than the LASSO. This is the case in our study of interpreting property price using amenities and property characteristics.

Model Interpretability

As highlighted by previous studies, machine learning approaches are effective in selecting regressors or features (Guyon, 2003). For LASSO Model (2a), different thresholds for λ are set to compare the relative impact of regressors on the estimates of β. For higher-order terms of MRT distance d_{MRT}, LASSO favors its cube term over d_{MRT} and its square term. However, replacing the first and second order terms in OLS by the cube term does not show a significant improvement in the goodness-of-fit. Among all interactions, LASSO selects $d_{MRT} * age$ to be the most significant term. We find that including this interaction term will improve the coefficient of determination of the model.

Using RF approach, we calculate Gini importance, i.e., the variable importance levels based on determinants' predictive power to increase prediction accuracy (how much including a variable increases prediction accuracy). This feature importance can be applied to all tree-based learning, including, decision trees, boosting, and RF. An alternative approach involves out-of-bag (OOB) errors to rank the variables. The OOB is a special feature that only exists in methods incorporating bagging tree or random forest. For a detailed description on both methods, see Hastie et al. (2009, p.593). The results from Model (3a) using Gini importance and OOB randomization importance are presented as the left plot and right plot in Figure 4. Since the index is a relative amount, the largest one is set to be 100.

The most important determinant of property prices is *Time* (month dummies), which captures the macroeconomic factors affecting both Singapore economy as well as the real estate market at the point of transaction. *Planning_area* of the property is also greatly significant in determining property prices. Since planning_area is used to describe the location characteristics of the property. Beyond these, the distances to nearest MRT station is the 4th most important factor in determining property prices. It shows that buyers value the ease of access to MRT transportation highly when purchasing a property.

[1] The Wilcoxon signed-rank test proposed by Diebold and Mariano (1995) is conducted to compare the forecast accuracy of OLS with LASSO, RF and ANN algorithms. All p-values of these tests are less than 1%, suggesting that the null of a zero-median loss differential is rejected.

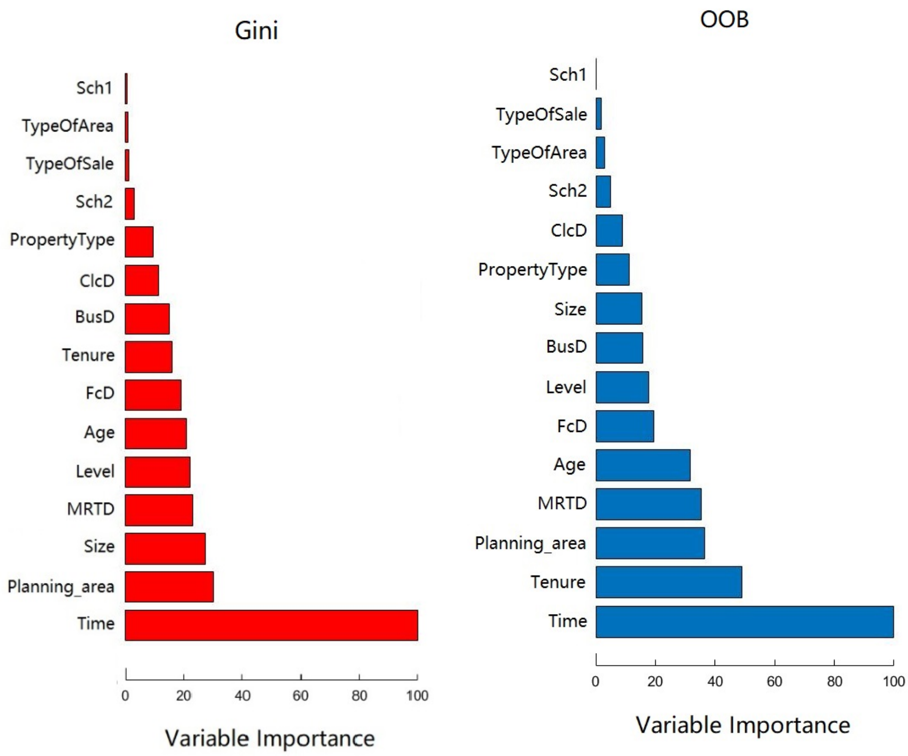

Figure 4 Variable Importance Plots

Due to the complexity of the ANN model, we are unable to obtain a single coefficient that captures the impact of distance to the nearest MRT station on property prices. Instead, we use a set of randomized control trials to measure the change in property prices by changing the distances to the nearest MRT station from its mean value and the weighted average of MDP derived during randomized control trials. The results from the ANN approach indicate that moving a property by 100 meters closer from the mean point, we observe an average increase in property price by 9,281 SGD (vs. 15,066 SGD using OLS).

5.3 A Comparison between Econometric Analyses and Machine Learning Approaches

In Sections 4 and 5, we estimate the effects of amenities and important factors on property price with a focus on the premium of MRT distance in the context of Singapore's private property market, using both an econometric approach and various machine learning algorithms. Both approaches have shown effective in delivering interesting results in this specific context.

Firstly, the econometric analyses based on a hedonic model give robust estimates of effects of amenities on property prices, including MDP in various scenarios. The adjusted R^2 of these specifications are high, suggesting that the relationship between the property price and regressors used in Table 3 are highly linear. The easy interpretation of OLS coefficients makes the OLS a popular and widely used method, including in constructing the official property price index.

Secondly, compared with the OLS, machine learning approaches, including LASSO, RF and ANN, provide new tools and deliver different results. The LASSO approach can be utilized to help select nonlinear forms of covariates, i.e., the interaction terms in this example. The RF approach used in this study shows a significant improvement in prediction accuracy. Moreover, a list of variable importance produced by the RF approach in Figure 4 provides a new insight on the relationship between the response and determinants. Similar to the RF approach, the ANN algorithm improves prediction accuracy and captures heterogenous effects by using randomized control trials. However, since the economic meaning behind the hidden layers and hidden units is unclear, and relatively large datasets are required to compensate for the loss of degree of freedom.

Comparing the econometric analyses and machine learning approaches in this empirical context, we find that they are two different methodologies dealing with different research questions. They have their own strengths and limitations in general. In a well specified empirical scenario supported by organized data sets, an econometric approach, e.g., OLS is preferred. With the help of economic theory, e.g., hedonic model in a housing market with detailed transaction data, the regression setup is well specified, and parameters have structural interpretation. Thus, it is easy to be implemented and interpreted. In this case, machine learning algorithms may help add additional insights, e.g., producing high prediction accuracy, but contribute little to the causal inference based on econometric analysis.

However, in an unstructured scenario with relatively raw data, machine learning approaches could be more useful than the OLS or other econometric approaches since their advantage is on prediction in a flexible and data-driven fashion. Thus, these new approaches could help find certain correlation and association among variables in data for further economic analysis. Thus, machine learning algorithms contribute to a more comprehensive understanding of the datasets and research problems. The effectiveness to measure existing association presented in dataset is proven by seeing an improvement in prediction accuracy. In addition, variable importance measured in the model training process provides a systematic feature selection reference for researchers to evaluate more complex relationship.

6. CONCLUSION

Using Singapore private property transaction data from 1995 to 2018, this paper quantifies premiums of amenities and important characteristics with a focus on the effect of distance to nearest MRT using econometric models and machine learning approaches. Regression estimates based on a hedonic model show that 1% decrease in distance to the nearest MRT station will increase the property price by 0.067% on average. This implies that moving a property 100 meters closer from the mean distance point to the nearest MRT station, we expect an increase in the overall transacted price by 15,066 SGD.

Meanwhile, we observe a 26.8% increase in the magnitude of MRT distance premiums after the global financial crisis in 2008. After the financial crisis, the Singapore government imposed strict limitations on mortgage loans to suppress speculative investments. This could result in a higher buyers' value on practical benefits, such as closeness to an MRT station. In addition, a scenario analysis suggests that the impact of distance to MRT stations on property prices diminishes beyond 700 meters, as MRT stations are no longer within a walking distance. This observation is in line with Diao et al. (2017), who show that with the opening of circle line, properties within 600 meters of new stations are the most heavily impacted. If more MRT lines are built, we should expect a spike in prices of properties that lie within a 700-meter radius of new stations.

Finally, 3 machine learning algorithms are also applied to this empirical context. In general, these algorithms produce better predict accuracy and provide flexible relationship between property price and amenities and characteristics, thus adding new sights to regression results. In this specific context, we compare econometric analyses with machine learning approaches and point out their pros and cons. Generally speaking, they are two different methodologies designed to address different research questions. In an empirical scenario assisted with economic theory and well-organized data sets, an econometric approach is preferred. However, in a scenario with an unstructured data, machine learning approaches could be useful to identify certain pattern on correlation among underlying variables in the data for further economic analysis.

7. REFERENCES

Au, TC (2018). Random forests, decision trees, and categorical predictors: the "absent levels" problem. *Journal of Machine Learning Research*. 2018, 19(45): 1–30.

Bajic, V. (1983). The Effects of a New Subway Line on Housing Prices in Metropolitan Toronto. *Urban Studies*, 20(2), 147-158.

Bency, A. J., Rallapalli, S., Ganti, R. K., Srivatsa, M., & Manjunath, B. S. (2017). Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Beutel, J., List, S., and G von Schweinitz, (2019) Does machine learning help us predict banking crises? *Journal of Financial Stability*, 45, 100693.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Bussiere, M., and M. Fratzscher. 2006. Towards a new early warning system of financial crises. *Journal of International Money and Finance* 25: 953-73.

Chan L, Ng HT, Ramchand R. (2012). A cluster analysis approach to examining Singapore's property market and financial stability. *BIS Pap Bank Int Settlements* 64:43–53

Debrezion, G., Pels, E., & Rietveld, P. (2011). The impact of rail transport on real estate prices: an empirical analysis of the Dutch housing market. *Urban Studies*, 48, 997–1015.

Deng, Y., Gyourko, J., & Li, T. (2019). Singapore's Cooling Measures and Its Housing Market. *Journal of Housing Economics*. Vol. 45, 101573.

Department of Statistics Singapore. (2015). Average Monthly Household Income by Working Status/Occupation of Main Income Earner and Type of Dwelling.

Diao, M., McMillen, D., & Sing, T. (2017). A Quantile Regression Analysis of Housing Price Distributions Near MRT Stations.

Diao, M., Fan, Y., & Sing, T. (2017). A New Mass Rapid Transit (MRT) Line Construction and Housing Wealth: Evidence from the Circle Line. *Journal of Infrastructure, Policy and Development*, 1(1): 64-89.

Diebold, F.X., Mariano, R.S. (1995), Comparing predictive accuracy, *Journal of Business & Economic Statistics*, 13(3), 253-263.

Ferrari, S., and Stengel, R.F. (2015). Smooth function approximation using neural networks. *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 24–38

Fesselmeyer, E., and H. Liu. (2018). How Much Do Users Value a Network Expansion? Evidence from 22 the Public Transit System in Singapore. *Regional Science and Urban Economics*, 71: 46-61.

Guyon, I., and A. Elisseeff, A. (2003) An introduction to variable and feature selection. *J. Mach Learn Res.*, 3, 1157–1182.

Hastie, T., R. Tibshirani, and J. Friedman (2009): The Elements of Statistical Learning, *Springer Series in Statistics*, New York, NY, USA.

Hall, D., Hirsch, P and Li, T. M. (2011). Powers of exclusion: land dilemmas in Southeast Asia, Singapore: NUS Press

Hu, L., He, S., Han, Z. & Xiao, H. & Su, S. & Weng, M. & Cai, Z. (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy*. 82. 657-673.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Kim, K., & Lahr, M. L. (2014). The impact of Hudson–Bergen light rail on residential property appreciation. *Papers in Regional Science*, 93, S79–S97.

Kang, J. S., P. Kuznetsova, M. Luca, and Y. Choi. 2013. "Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews." EMNLP 2013: 2013 *Conference on Empirical Methods in Natural Language.*

Lancaster, K.J. A New Approach to Consumer Theory. *Journal of Political Economy*. 1966, 74, 132–157

Li, Y., Jiao, C., Xin, H., Wang, Y., Wang, K. (2018). Prediction on Housing Price Based on Deep Learning. *International Journal of Computer and Information Engineering*, Vol:12, No:2.

Louppe, G., 2014, Understanding random forest: from theory to practice. PhD dissertation, University of Liège.

Lu, Y., Zhou, Y., Qu, W., Deng, M., & Zhang, C. (2011). A lasso regression model for the construction of microrna-target regulatory networks. *Bioinformatics* 27 (17), 2406–2413

Mayor, K., Lyons, S., Duffy, D., & Tol, R. S. J. (2008). A hedonic analysis of the value of rail transport in the greater dublin area. *Journal of Transport, Economics and Policy*, 46, 239–261.

McMillen, D. P., & McDonald, J. (2004). Reaction of house prices to a new rapid transit line: Chicago's midway line, 1983–1999. *Real Estate Economics*, 32, 463–486.

Melkumova, L.E., & Shatskikh, S. Ya. (2017) Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, Vol 201, 746-755.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), 87-106.

Pan, Q., Pan, H., Zhang, M., & Zhong, B. (2014). Effects of rail transit on residential property values: Comparison study on the rail transit lines in Houston, Texas, and Shanghai, China. *Transportation Research Record*, 2453(1), 118–127.

Rosen, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 1974, 82, 34–55

Sobolevsky, S. & Sitko, I. & Tachet des Combes, R. & Hawelka, B. & Arias, J. & Ratti, C. (2014). Money on the Move: Big Data of Bank Card Transactions as the New Proxy for Human Mobility Patterns and Regional Delineation. The Case of Residents and Foreign Visitors in Spain. *Proceedings - 2014 IEEE International Congress on Big Data,* 136-143.

Tanaka, K., Higashide, T., Kinkyo, T., and Hamori, S. (2019) Analyzing industry-level vulnerability by predicting financial bankruptcy, *Economic Inquiry*, 57 (4), 2017-2034.

Tanaka, K., Kinkyo, T., & Hamori, S. (2016). Random forests-based early warning system for bank failures. *Economics Letters*, 148, 118–121.

The Economist. (2017). Why 80% of Singaporeans live in government-built flats. July 6th 2017 edition.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B,* 58(1), 267-288.

Wang, S., and Zhao, J. (2017). Using deep neural network to analyze travel mode choice with interpretable economic information: An empirical example, arXiv:1812.04528.

Table 1 Distribution of Transactions By Distance Range to MRT Stations

Distance Range	Number of Transactions	Percentage
0-200m	52,143	11.2%
200-500m	173,197	37.1%
500-1,000m	181,854	39.0%
1,000-2,000m	58,556	12.5%
2,000m and above	867	0.2%
Total	466,617	100.0%

Data source: REALIS

Table 2 Descriptive Statistics of Variables

Symbol	Definition	Unit	Mean	Std. D.	Form in regression	Data sources
y	Property price per square meter	1 SGD	10,441	5334	log	REALIS
area	Total area	1 square meter	130.0	97.8	log	REALIS
d_{MRT}	Distance to the nearest MRT	1 meter	603.6	398.9	log	OneMap SG
d_{bus}	Distance to the nearest bus stop	1 meter	153.5	98.0	log	OneMap SG
d_{food}	Distance to the nearest food court	1 meter	954.5	643.4	log	OneMap SG
d_{clinic}	Distance to the nearest clinic	1 meter	288.9	200.4	log	OneMap SG
age	Age of the property	1 year	4.1	7.1	discrete	REALIS, Authors' Calculation
level	Floor level	1 floor	8.4	7.6	discrete	REALIS
SCHOOL1	With at least one top primary school within 1km				dummy	OneMap SG, MOE
SCHOOL2	With at least one top primary school between 1-2km				dummy	OneMap SG, MOE
TENURE	Tenure level (99-years leasehold, 999-years leasehold, freehold)				dummy	REALIS, Authors' Calculation
DIST	Planning area (38 districts)				dummy	REALIS
LAND	Type of area (landed vs strata)				dummy	REALIS
TYPE	Type of properties (apartment, detached house, etc)				dummy	REALIS
SALE	Type of sale (new sale, resale and subsale)				dummy	REALIS
MONTH	month of the transactions				dummy	REALIS

Note: Units and summary statistics of all variables are reported before taking log.

Table 3 Effects of Amenities on Property Prices

Dependent variable: property price per square meter

Indep. Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
d_{MRT}	-0.067***	-0.026***	-0.077***	-0.056***	-0.071***	-0.090***	-0.040***	-0.010***	-0.067***
	(0.001)	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)
d_{bus}	0.007***	0.007***	0.008***	0.010***	0.005***	0.018***	-0.001*	0.002	0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
d_{food}	0.035***	0.035***	0.035***	0.061***	0.009***	0.042***	0.009***	0.016***	0.030***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)
d_{clinic}	0.015***	0.015***	0.015***	0.032***	0.007***	0.026***	-0.0016***	0.046***	0.013***
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.003)	(0.000)
SCHOOL 1	0.006***	0.006***	0.006***	0.026***	-0.005**	0.003*	0.018***	0.019***	-0.000
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.003)	(0.001)
SCHOOL 2	0.009***	0.009***	0.008***	0.030***	-0.006***	-0.007***	0.034***	0.080***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)
d_{MRT}, *squared*	-	-0.004***	-	-	-	-	-	-	-
	-	(0.000)	-	-	-	-	-	-	-
d_{MRT} * *age*	-	-	0.003***	-	-	-	-	-	-
	-	-	(0.000)	-	-	-	-	-	-
	All	All	All	1995-2008	2009-2018	Central	non-central	Land	Strata
Type of area	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sale type	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Property type	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tenure	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Planning area	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Month dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.87	0.87	0.87	0.79	0.84	0.84	0.89	0.83	0.88
No. of obs.	466,617	466,617	466,617	205,293	261,324	207,750	258,867	35,636	430,981

Note:

1. Standard errors are reported in parentheses. The symbols *, ** and *** indicate the significance level at 10%, 5%, 1% and 0.1%, respectively.
2. For the definition, unit of variables and data sources, please refer to Table 2.

Table 4 Effects of Amenities on Property Prices: Additional Robustness Check

Dependent variable: property price per square meter

	Trim upper and lower 5%		
Indep. Var.	(1)	(2)	(3)
d_{MRT}	-0.0673***	-0.0263***	-0.0769***
	(0.0005)	(0.0042)	(0.0006)
d_{bus}	0.0069***	0.0067***	0.0074***
	(0.0005)	(0.0004)	(0.0005)
d_{food}	0.0341***	0.0345***	0.0339***
	(0.0006)	(0.0006)	(0.0006)
d_{clinic}	0.0151***	0.0149***	0.0145***
	(0.0004)	(0.0004)	(0.0004)
$SCHOOL\,1$	0.0065***	0.0062***	0.0066***
	(0.0009)	(0.0009)	(0.0009)
$SCHOOL\,2$	-0.1555***	0.0082***	0.0078***
	(0.0008)	(0.0007)	(0.0007)
d_{MRT}, *squared*	-	-0.0035***	-
	-	(0.0004)	-
d_{MRT} * *age*	-	-	0.0025***
	-	-	(0.0006)
Type of area	Yes	Yes	Yes
Sale type	Yes	Yes	Yes
Property type	Yes	Yes	Yes
Tenure	Yes	Yes	Yes
Planning area	Yes	Yes	Yes
Month dummy	Yes	Yes	Yes
Overall Adjusted R^2	0.87	0.87	0.87
No. of observations	419,956	419,956	419,956

Notes:

1. Standard errors are reported in parentheses. The symbols *, ** and *** indicate the significance level at 10%, 5%, 1% and 0.1%, respectively.
2. For the definition, unit of variables and data sources, please refer to Table 2.

Table 5 Performance of Different Algorithms in Private Property Unit Price Prediction

Methodology	OLS		LASSO				RF				ANN	
Model	(1)		(2)		(2a)		(3)		(3a)		(4)	
Performance	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2	MSE	R2
Training Set	0.032	86.2%	0.032	86.2%	0.024	89.2%	0.003	98.7%	0.001	99.6%	0.020	91.4%
Test Set	0.032	86.1%	0.032	86.0%	0.023	89.3%	0.021	90.7%	0.005	97.7%	0.021	91.2%
Gain over OLS	-	-	0.0%	0.0%	-27.6%	3.8%	-33.8%	5.4%	-83.4%	13.5%	-36.8%	6.0%

Note: For model specifications, please refer to Section 4.

Table 6 MSE and R^2 with Number of Trees (N) in RF

Model		(3)		(3a)	
N	Dataset	MSE	R^2	MSE	R^2
100 trees	train	0.0033	0.986	0.0009	0.996
	test	0.0225	0.903	0.0056	0.976
200 trees	train	0.0032	0.986	0.0009	0.996
	test	0.0223	0.904	0.0056	0.976
300 trees	train	0.0032	0.986	0.0009	0.996
	test	0.0222	0.904	0.0056	0.976
400 trees	train	0.0032	0.986	0.0009	0.996
	test	0.0222	0.904	0.0056	0.976
500 trees	train	-	-	0.0009	0.996
	test	-	-	0.0056	0.976

Appendix 1 Definitions of Housing Attributes in the Dataset

Name	Definition
Project Name	String, name of the property in CAPS
Address	String, address of the property, mostly consist of the street name and number, as well as the unit number of the property
No. of Units	Integer, the number of units brought by the buyer
Area sqm	Integer, the size of the property in square meters
Type of Area	Classification, either Strata or Landed
Unit Price psm	Integer, the price per square meter in SGD
Unit Price psf	Integer, the price per square feet in SGD
Property Type	Classification, either Condominium, Apartment, Executive Condominium, Terrace House or Semi-detached House
Completion.Year	Integer, year of completion of the project, either an interger, uncompleted or unknow
Type of Sale	Classification, either New Sale, Resale or Sub-sale
Purchaser Address Indicator	Classification, either HDB, Private or N.A
Postal District	Integer, property falls under one of 28 postal districts in Singapore
Postal Sector	Integer, property falls under one of 82 postal sectors in Singapore
Postal Code	Integer, postal code of the property, as administered by Singapore Post
Planning Region	String, property falls under one of five planning regions in Singapore
Planning Area	String, property falls under one of fifty-five planning regions in Singapore
Tenure	Classification, Freehold, 999-year Leasehold and 99-years Leasehold are mostly seen
Trasaction Date	Date, Date of the transaction

Appendix 2 Singapore Top 21 Primary School List Based on Received MOE Awards

School Name
Raffles Girls Primary School
Rulang Primary School
St. Hildas Primary School
Ai Tong School
Catholic High School
CHIJ St. Nicholas Girls School
Chongfu School
Gongshang Primary School
Henry Park Primary School
Kong Hwa Primary School
Kuo Chuan Presbyterian Primary School
Maris Stella High School
Methodist Girls School
Nanyang Primary School
Pasir Ris Primary School
Pei Chun Public School
Radin Mas Primary School
Rosyth School
Tampines Primary School
Tao Nan School
Temasek Primary School

Note: MOE awards include School Excellence Award (SEA), School Distinction Award (SDA), Best Practice Award (BPA), Outstanding Development Award (ODA) and Development Award (DA). The number of awards received in the past 5 years
are counted.

Appendix 3 Number of Transactions within Each Planning Area

Planning Area	Number of Transactions	Percentage
Ang Mo Kio	11,300	2.42%
Bedok	44,961	9.64%
Bishan	9,756	2.09%
Bukit Batok	19,259	4.13%
Bukit Merah	15,405	3.30%
Bukit Panjang	9,442	2.02%
Bukit Timah	28,921	6.20%
Changi	51	0.01%
Choa Chu Kang	11,597	2.49%
Clementi	14,081	3.02%
Downtown Core	8,678	1.86%
Geylang	18,489	3.96%
Hougang	21,764	4.66%
Jurong East	3,821	0.82%
Jurong West	12,284	2.63%
Kallang	18,261	3.91%
Mandai	971	0.21%
Marine Parade	13,984	3.00%
Museum	1,248	0.27%
Newton	7,720	1.65%
Novena	19,614	4.20%
Orchard	1,496	0.32%
Outram	1,679	0.36%
Pasir Ris	23,387	5.01%
Punggol	8,380	1.80%
Queenstown	13,544	2.90%
River Valley	11,175	2.39%
Rochor	4,396	0.94%
Sembawang	5,039	1.08%
Sengkang	15,411	3.30%
Serangoon	18,945	4.06%
Singapore River	6,019	1.29%
Sungei Kadut	133	0.03%
Tampines	17,785	3.81%
Tanglin	15,490	3.32%
Toa Payoh	11,874	2.54%
Woodlands	9,511	2.04%
Yishun	10,746	2.30%
Total	466,617	100.00%
Average	12,279	2.63%
Min	51	0.01%
Max	44,961	9.64%

Note: According to REALIS, Outram, Museum, Newton, River Valley, Singapore River, Marina South, Marina East, Straits View, Rochor, Orchard and Downtown Core are considered to be in the central region.