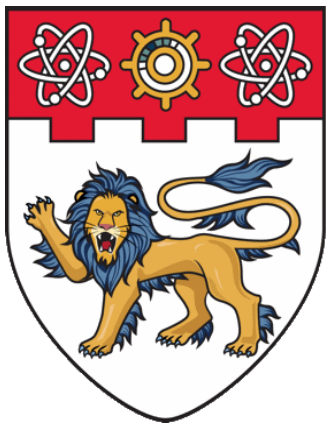
GP Project Number: HE1AY2021-08

NANYANG TECHNOLOGICAL UNIVERSITY

SCHOOL OF SOCIAL SCIENCES



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

PRICE DETERMINANTS OF PUBLIC HOUSING IN SINGAPORE: AN ECONOMETRIC AND MACHINE LEARNING APPROACH

Written by:

Zachary Tang Jia Ying	U1731121D
Benedict Peh	U1730325C
Daren Lim Jun Hao	U1730041L

A Graduation Project submitted to the School of Social Sciences, Nanyang Technological University in partial fulfillment of the requirements for the Degree of Bachelor of Arts in Economics

Academic Year: 2020/2021

ABSTRACT

Understanding the price determinants of real estate and the ability to predict prices can be beneficial for homebuyers, homeowners and policymakers. This paper seeks to examine the price determinants in Singapore's public housing market using both econometrics and machine learning approaches. The paper also studies feature importance within machine learning models to understand which variables are the most significant in improving predictive performance. Using transactional data from 2018-2020, we estimated a hedonic pricing model which quantifies the price effects of important housing attributes with a focus on a set of 18 locational amenities. Results showed a large and significant price premium for the proximity to a MRT (Mass Rapid Transport) station. In particular, an increase in proximity by 1 kilometre to a MRT station will increase flat prices by S$36,843 on average. The price effect is larger compared to bus interchanges and bus stops, suggesting that consumers value MRTs over bus transportation. In terms of education, we find that the proximity to a top-ranking primary and secondary school had a price premium of S$2,532 and S$13,999 on average respectively, significantly outweighing the price decreases due to noise and congestion from the schools. Results also indicate a S$15,601 price discount on average for every kilometre increase in proximity to nursing homes, quantifying the negative perceptions that consumers have of nursing homes. For machine learning, seven algorithms (LASSO, RIDGE, K-Nearest Neighbours, Artificial Neural Network, Regression Tree, Random Forest, Boosted Trees) were trained and evaluated based on the out-of-sample prediction accuracy. Our findings showed that nonlinear models outperformed linear models, with the best prediction accuracy produced by Artificial Neural Networks. In terms of feature importance, we find that the proximity to MRT stations and government hawkers were important considerations for homebuyers. This finding is robust across different machine learning methods. For other amenities, the findings were less conclusive due to their low relative importance.

TABLE OF CONTENTS

INTRODUCTION

As of 2019, 78.7% of total Singapore resident households live in public housing apartments, known as HDB flats, built and sold by the Housing Development Board (HDB) of Singapore (Singstat, 2019). New HDB flats are sold in the heavily state-controlled and demand-led primary market, which has resulted in long waiting times for new flats (Phang & Helble, 2016). Those who do not wish to wait can buy resale HDB flats in the secondary market where prices are determined by market forces.

Understanding the price determinants of resale HDB flats and the ability to predict prices can be beneficial to homebuyers, homeowners and policymakers. Homebuyers would be able to make informed decisions in home purchases. Homeowners looking to sell would be able to accurately value their houses. Policymakers would be able to make more informed decisions for housing affordability. However, the level of heterogeneity between houses makes identifying the determinants of prices difficult. Traditional models rely on hedonic price regressions (Wolfgang & Steininger, 2020) but they are limited by their explicit assumptions. There has been a growing amount of literature deploying Machine Learning (ML) approaches in real estate. Instead of estimating parameters for causal inference, ML's goal is to maximize prediction accuracy. A key advantage of using ML over econometric models is the data-driven nature of ML through the use of algorithms (Athey, 2019), which tends to yield more accurate predictions.

Despite the extensive literature studying house price determinants, there has been a lack of research in Singapore's public housing market. In this paper, we first aim to study resale flat price determinants using econometric and ML approaches. Secondly, we investigate feature importances from selected ML models to derive additional insights. We hope that this paper can enable policymakers, homeowners and homebuyers to make better-informed decisions in the public housing market.

The paper is organized as follows. Section 2 reviews the relevant literature. Section 3 provides an overview of Singapore's public housing sector. Section 4 covers the methodology of our research, including the data collection process, econometric and ML methodologies. Section 5 presents the empirical results and limitations of the study. Section 6 concludes the paper.

LITERATURE REVIEW

There exists a large number of literature studying the characteristics that determine house prices. To do so, researchers have mostly relied on the hedonic price model based on the consumer theory presented by (Lancaster, 1966) and further built upon by (Rosen, 1974). Based on the model proposed by (Rosen, 1974), house prices can be viewed as a set of characteristics that the consumer values, such as house location, structure, neighbourhood and environmental amenities. Studies using hedonic regression by (Wen et al., 2005), (Heyman & Sommervoll, 2019), (Munoz-Raskin, 2010) found significant effects of amenity accessibility on house prices. Research by (Efthymiou & Antoniou, 2013), (Mulley & Tsai, 2016), (Bae et al., 2003), (Bowes & Ihlanfeldt, 2001), (Cao & Hough, 2012) in various countries show significant effects of transport proximity on house prices with mixed signs, implying that there exist different perceptions of value towards transportation amenities across countries. On educational institutions, (Yi et al., 2017) demonstrated a positive relationship between school performance and proximate house prices in Seoul. Studies by (Chiodo et al., 2010) and (Seo & Simons, 2009) show similar results. For other amenities, (McCord et al., 2014) and (Wu et al., 2015) found a positive effect of park proximity on house prices. While transport, education and recreation locations are intensely studied in existing literature, we note a lack of studies on the impact of healthcare facilities and food amenities on house prices.

In Singapore's context, (Sue & Wong, 2010) utilised a hedonic pricing model and found positive price effects attributed to the proximity to top schools and Mass Rapid Transport (MRT) stations, however, the study was limited to only 2 constituencies. Studying private properties, (Addae-Duppah & Lan, 2010) found a positive effect of shopping centres on proximate houses. (Bian et al., 2018) also found that moving a property 100 metres closer to a MRT station resulted in a S$15,000 increase in price. However, these studies were limited by only considering a handful of locational amenities.

Next, we focused on literature applying machine learning methodologies in real estate. (Baldominos et al., 2018) trained several ML models using online house listing data in Salamanca, Spain and they found that K-Nearest Neighbors produced the most accurate results. (Ho et al., 2021) found that the predictions from Support Vector Machines were the most accurate using 39,554 housing transactions in Hong Kong. (Huang, 2019) reported that linear regression produced the most accurate predictions after training several linear and non-linear ML algorithms using 89,412 housing transactions in Los Angeles, California. These studies were limited by only considering the physical features of the houses. Using

both physical and locational features with 139,954 housing transactions in Australia, (Gao et al., 2019) found that a multi-task learning algorithm produced the best predictions.

In Singapore's context, (Bian et al., 2018) reported the best prediction accuracy using the Random Forest algorithm with a set of 516,962 private property transactions and locational features. The authors also studied feature importances within the algorithm and found that the time of the transaction was the most important feature for prediction. (Wang et al., 2016) trained a Deep Neural Network on online HDB listing data with locational and macroeconomic features. The authors found that the neural network was effective in mapping nonlinear relationships between flat prices and housing characteristics and thus produced good prediction results. However, these studies were limited by the number of models trained.

While there is substantial research studying house price determinants in both econometric and ML literature, there is a lack of research in the Singapore context. The limited studies in Singapore are also narrow in terms of the number of locational variables in the data and the variety of models estimated. Our study hopes to build upon the literature by including locational amenities such as healthcare and food establishments and evaluating a larger number of models for insight.

SINGAPORE PUBLIC HOUSING MARKET

3.1 Overview

HDB was founded in 1960 to provide housing for the poor to rent. This has since evolved to the development of housing estates aimed at making housing affordable for all citizens in Singapore. Due to land scarcity, HDB flats are returned to the government upon expiry of a 99-year lease.

3.2 Market of HDB Resale Flats

Resale HDB flats are sold in the secondary housing market. Unlike the prices of new flats which are determined by HDB, resale flat prices can fluctuate due to market forces.

FIGURE 1: RESALE HDB PRICE INDEX FROM 1990 - 2020

Since 1990, flat prices have been increasing, as seen in Figure 1. To cool the market, the government introduced a slew of cooling measures since 2013. Despite the measures, there has been a large number of resale transactions since 2013 (Lim & Seah, 2017), which brought up questions about the effectiveness of the measures.

3.3 Major Policy Changes from 2018 to 2020

In 2018, the Voluntary Early Redevelopment Scheme was introduced to allow the government to buy back flats that are 70 years and older for redevelopment. This allows homeowners a chance to give up their flats early to the government through resident voting. Furthermore, the government announced various cooling measures by increasing the Additional Buyer's Stamp Duty and tightening the Loan-to-Value limits for residential properties. The measures were aimed to deter Singaporeans from owning multiple properties.

In 2019, HDB replaced the Additional Housing Grant and Special Housing Grant with a single Enhanced Housing Grant (EHG) where first-time home purchasers can receive up to S$80,000 in grants depending on their income. EHG can be applied to all flat types, unlike the previous grants where the income ceilings were tiered according to the flat type.

In March 2020, the government allowed unwed parents to purchase 3-room flats from HDB. Previously, these individuals could only purchase 2-room flats from non-mature estates. However, they are still not eligible for the Families Grant scheme which provides additional subsidies for families.

DATA

4.1 Data Collection

Data used in this study was obtained from a variety of online sources.

4.1.1 Resale HDB Transaction Data

67,902 resale transaction records between January 2018 - December 2020 were extracted from data.gov.sg, an online government portal for publicly-available datasets. Variables in the transaction data include the transaction month and year, address, resale price, floor area in square metres, storey range, flat model, flat type and lease commencement date. Remaining lease was calculated by first taking the difference between the lease commencement date and transaction time, then taking the difference again with the 99-year lease. The time range of 2018-2020 was chosen as we were not able to ascertain if all locational variables existed in the same period. In this paper, we assume that all locational features gathered existed between 2018 and 2020.

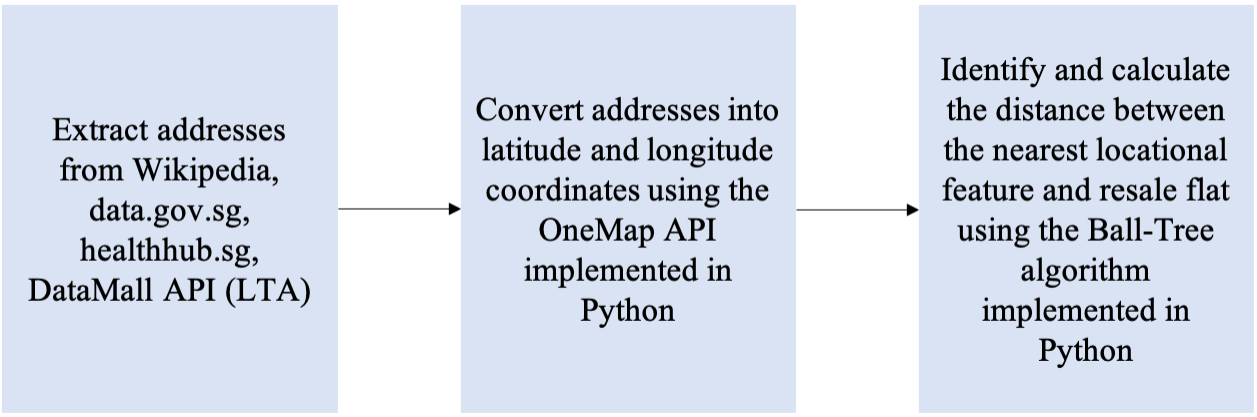
4.1.2 Locational Features

To build upon the paper by (Bian et al., 2018), we included 18 locational features (compared to 6 in the author's paper) into our study to capture the price effects of amenities. Locational features were categorized into 4 groups: transport, education, recreation and health. In the transport group, the locational features include the distance to the nearest Mass Rapid Transport (MRT) station, bus station (interchange) and bus stop. Education group features included the distance to the nearest primary school, secondary school and junior college. We also included dummy variables to indicate if the resale HDB flat is within a 1 kilometre radius of a top 25 primary school, top 25 secondary school and top 5 junior colleges. MOE does not officially publish school rankings, hence we derived the ranking by calculating the number of MOE education awards received in the past five years. Recreation group features included the distance to the nearest shopping centre, government hawker, National Environment Agency (NEA) licensed food stall, supermarkets, child care centres, parks and

community clubs. Lastly, the Health group features included the distance to the nearest medical clinic, nursing home and retail pharmacy.

Figure 2 provides a brief overview of the data gathering process for locational features.

FIGURE 2: LOCATIONAL FEATURE DATA GATHERING PROCESS



For the Transport group of locational features, a list of all MRT and bus interchange addresses in Singapore was extracted from Wikipedia. For bus stops, we extracted the addresses from the DataMall API, which is an API maintained by the Land Transport Authority (LTA) of Singapore to provide users with transport-related datasets. In the Education group, we collected the names and addresses of all primary schools, secondary schools and junior colleges from their respective Wikipedia pages. For the Recreation group, we extracted the addresses from data.gov.sg. Lastly, for the Health group, we extracted the addresses from Healthhub.sg, a one-stop online portal for health-related content and e-services for Singaporeans, maintained by the Ministry of Health.

The OneMap API, a national map of Singapore maintained by the Singapore Land Authority, was implemented in Python to convert the addresses into latitude and longitude coordinates. Table 1 in the appendix provides a summary of all unique locations generated for this study.

Lastly, to identify the nearest locational feature and calculate the distance between the feature and the resale flat, we implemented the Ball-Tree algorithm in Python. The algorithm organizes data points into "ball trees", a space-partitioning data structure that is constructed based on the great-circle distance between data points. The Haversine formula, commonly used in navigational applications, was used to calculate the great-circle distance.

4.1.3 Categorical Data

Using the OneMap API, we were able to generate the planning area of the flats using their respective latitude and longitude coordinates. Planning areas are the main urban planning and census divisions of Singapore as delineated by the Urban Redevelopment Authority of Singapore. There are a total of 55 planning areas, however, only 32 planning areas contained HDB flats. Each planning area captures the demographic factors in that area.

Finally, the categorical data in our dataset was transformed into dummy variables using one-hot encoding and ordinal encoding. One-hot encoding transforms each category in the categorical variable into a unique dummy variable. Ordinal encoding transforms the categorical variable into a set of discrete ranked values. In the case of our study, variables flat type and storey range were converted into dummies using ordinal encoding as these variables had an innate ranking associated with them. Variables planning area, flat model, transaction month and transaction year were converted to a set of dummy variables using one-hot encoding.

The final dataset used in this study consists of 67,092 observations and 87 independent variables. Table 2 in the appendix provides a summary of variables in the dataset.

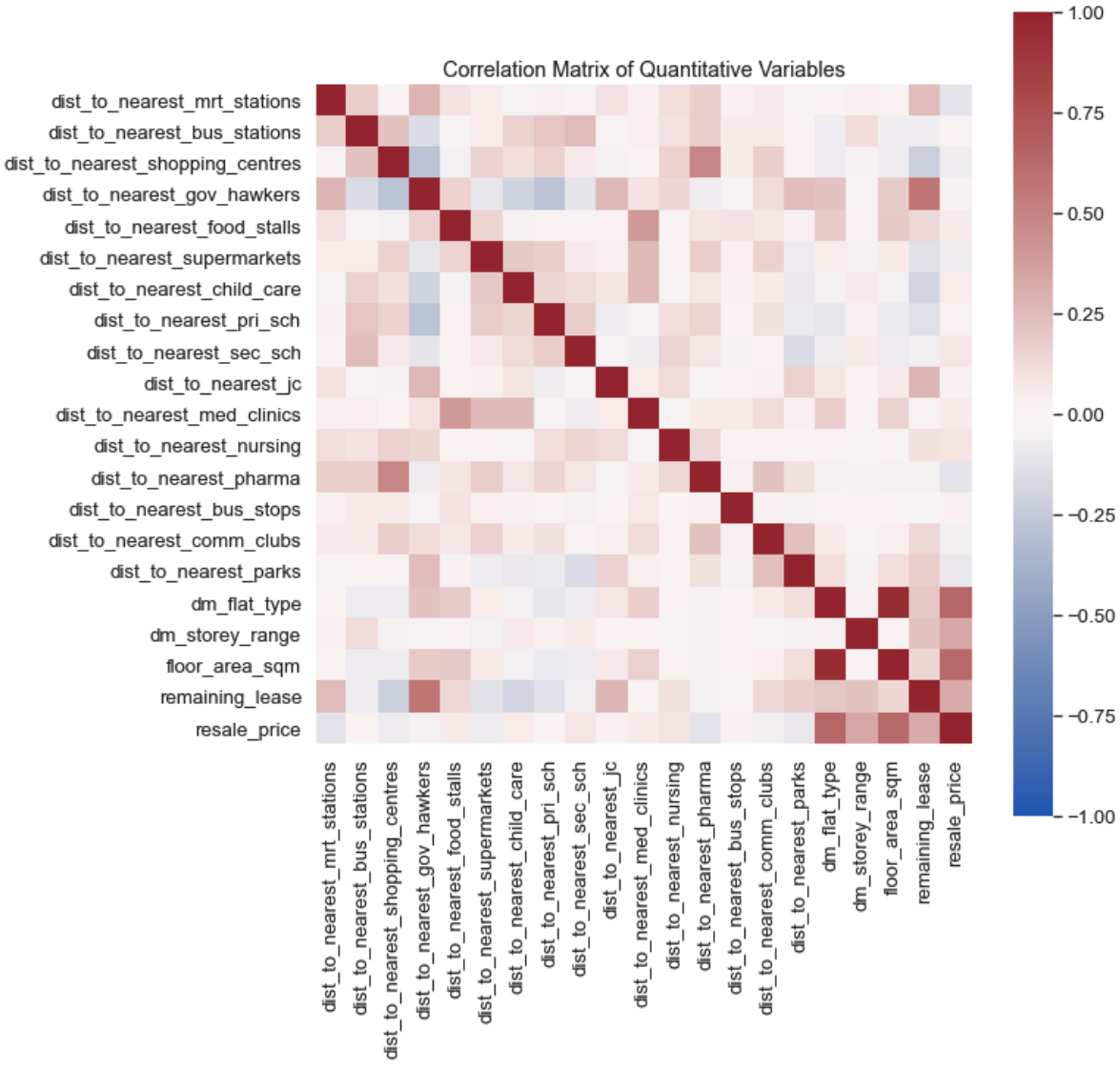
4.2 Exploratory Data Analysis

Table 3 in the appendix shows the descriptive statistics of all quantitative variables in our dataset. An interesting observation is the low mean distances to the nearest bus stop, food stall and child care centre suggesting that these facilities can be commonly found in housing estates.

Figure 3 in the appendix presents a map of all resale HDB flat transactions in the data. Looking at the figure in the top left corner, we can see that resale flats in planning areas nearer to the central region of Singapore command a higher price. We should expect the planning area variables to have some predictive power in predicting resale prices. Looking at the plot in the bottom left corner, we see that resale flats sold in the northeastern region of Singapore, especially in the Sengkang and Punggol planning areas, have a higher remaining lease. This observation tells us that many HDB owners are selling their flats after meeting the Minimum Occupation Period of 5 years, perhaps to upgrade to private residences. Furthermore, these are relatively new estates and there might be a lack of amenities in the area.

Figure 4 presents a correlation matrix of all quantitative variables in the data. We find strong linear relationships between the resale price and variables flat type, floor area, storey range and remaining lease. We also note the positive correlations between the proximity of MRT stations and proximities for bus stations, government hawkers and retail pharmacies. This implies that these locations are close to each other. Similarly, the positive correlation between the distances to the nearest shopping centre and the distance to the nearest retail pharmacy indicates that pharmacies are usually found in shopping centres.

FIGURE 4: CORRELATION MATRIX OF ALL QUANTITATIVE VARIABLES



METHODOLOGY

5.1 Econometric Model

5.1.1 Model Specification

We estimate a hedonic price model using Ordinary Least Squares (OLS) to examine the determinants of resale flat prices. The model is specified as follows:

$$y = c + \beta x_1 + \gamma x_2 + \delta x_3 + e$$

where y is the resale HDB flat price in thousands. x_1 denotes a vector of physical characteristics including floor area in square meters, remaining lease and flat type. x_2 denotes a vector of distance variables, capturing the distance in kilometres between the flat to the nearest supermarket, shopping centre, MRT station and so on. x_3 denotes a vector of controls, including flat model and time dummies. Time dummies capture the time-varying macroeconomic factors and policy changes. Finally, e represents the error term.

5.1.2 Model Assumptions

There is a need to assume that the hedonic model satisfies the Gauss-Markov assumptions for a causal interpretation. Linear in parameters is satisfied by the definition of the model. Random sampling is also satisfied as the resale transactions are random. Since the variables are not perfectly linear to another, no perfect collinearity is satisfied. For the assumption of zero conditional mean, it is worth noting that HDB flat specific factors like the condition of the flat or window facing directions are not controlled in the regression. We are unable to find strong evidence that these factors are related to the independent variables, hence we assume that the assumption holds. Our LM test results suggest that heteroskedasticity exists in the dataset. Therefore, we utilized heteroskedasticity-robust standard errors in the estimation of the hedonic model.

5.2 Machine Learning

In the following subsections, we will introduce key machine learning concepts and detailed methodologies for each model estimated in this paper.

5.2.1 Sample Splitting

In machine learning, data is split into two "training" and "test" subsamples. The "training" set is used to estimate the model and the "test" set is used as an out-of-sample dataset for which the model uses to produce a set of predicted outcomes. In this paper, the deviations between the predicted outcomes and the actual outcomes are measured by calculating the Root Mean Squared Error (RMSE).

In ML literature, there is no consensus on the optimal train-test split ratio. For our main analysis, we split our dataset according to a 4:1 ratio where 80% of the data is designated as the "training" set and the remaining 20% of the data is designated as the "test" set. The temporal ordering of the data was maintained, following the methodology in the literature.

5.2.2 Hyperparameter Tuning via Cross Validation

Hyperparameters are parameters in an ML model that cannot be estimated from the data and thus manually tuned in the model. This paper follows a common approach in literature known as K-fold cross validation for tuning hyperparameters. In cross validation, the training sample is split into a K-number of stratified subsamples (known as folds). K-1 folds are used to train the ML model and the Kth fold is used to calculate the cross validation score of the model. The process is repeated K times and the cross validation scores are averaged and recorded. Finally, hyperparameter values that resulted in the best cross validation score are selected.

Choosing the value of K is not straightforward due to the difficulty in estimating the variance for cross validation scores (Bengio & Grandvalet, 2004). Studies using real-world datasets show a 10-fold stratified cross validation works well even if computational power allows for a higher number of folds (Kohavi, 1995). As such, we will use a 10-fold cross validation in our analysis.

5.2.3 Bias-Variance Tradeoff

In supervised machine learning, prediction accuracy can be broken down to variance and bias components (James et al., 2013). High variance suggests that the model is overly-complex and overfits the data. High bias implies that the model is unable to capture the complex relationships between independent variables, which results in poor prediction accuracy. Increasing the complexity of the model reduces bias but introduces variance. Conversely, introducing simplicity increases bias and lowers variance. Hyperparameters are tuned to find a minimum balance between bias and variance.

5.2.4 LASSO

LASSO (Tibshirani, 1996) regression is similar to OLS regression but with a L1-norm penalty added to the regression coefficients. By penalizing the coefficients, unimportant variable coefficients may be reduced to zero and the result is a parsimonious model that minimizes overfitting. As a result of the penalty or "regularization", the estimates of the coefficients are biased, but the prediction accuracy on out-of-sample data is improved.

The hyperparameter λ controls for the amount of penalty imposed and is selected via cross validation.

5.2.5 RIDGE

RIDGE (Hoerl & Kennard, 1970) regression is similar to LASSO but instead of using an L1-norm penalty, RIDGE imposes a L2-norm penalty on regression coefficients. The effect of this penalty shrinks the coefficients towards zero but does not reduce them to zero. Similar to LASSO, the regularization introduces bias in the coefficient estimates as an expense for improved out-of-sample prediction accuracy.

The hyperparameter λ serves to control for the amount of penalty imposed and is selected via cross validation.

5.2.6 K-Nearest Neighbours

K-Nearest neighbours algorithm (KNN) is a non-parametric machine learning method used for classification and regression problems. Predictions from the KNN algorithm are based on the distances between in-sample data points and out-of-sample data points. Given a set of out-of-sample features, KNN algorithm first identifies K points in the in-sample data that are closest to each feature in the set, then estimates the value of the response variable as an average of the K nearest points.

Hyperparameters in this model include K, the number of neighbors to consider and D, the distance metric. K and D are chosen using cross validation.

5.2.7 Artificial Neural Network

Artificial Neural Networks (ANN) are nonlinear models designed to mimic the biological neural networks found in animal brains. ANNs can be thought of as universal approximators (Honik et al., 1989), with wide applications in speech recognition and autonomous vehicles. In this study, we focus

on “feed-forward” ANNs, where data moves through each layer in one direction. The model starts with the “input layer” of raw features and one or more “hidden layers” that are connected with the layer before it in a complex and nonlinear manner. Each layer consists of several nodes that are assigned a certain weightage. When data is passed through the node, an “activation” function is performed on the data before being passed to the next layer. All nodes converge in the output layer which produces the prediction.

Tuning an ANN is computationally expensive, as a result, we focused on tuning the number of hidden layers, nodes within each hidden layer, the activation function, number of iterations and the L1-norm regularization parameter.

5.2.8 Regression Tree (RT)

Regression tree is a non-parametric model that is “grown” in a sequence of steps. In the first step, the training data is split into two “branches” by dividing the data based on the most important feature and a split point that results in the largest decrease of the residual sum of squares. The process is then repeated with the leftover data until the set of all possible values for each feature has been allocated into a unique region known as “leaves”. A prediction is made by calculating the mean of the training observations in the “leaf” to which the test observation belongs (James et al., 2013). Regression trees are prone to overfitting since the tree is able to find a split for every data point in the feature set.

To control for overfitting, the tree is “pruned” or regularized by adding a cost complexity hyperparameter α, which penalizes the tree for having many splits or “leaves”. The maximum number of “leaves”, T, is also specified to reduce overfitting. Hyperparameters $\alpha$and T are chosen by cross validation.

5.2.9 Random Forest (RF)

Random Forest is an ensemble method that combines many different regression trees trained with a variation of the “bagging” method. “Bagging”, or bootstrap aggregating, is performed by repeatedly sampling from the training data with replacement and growing a regression tree on each bootstrap sample. Predictions are made by aggregating the predictions from all regression trees. Bagging has been shown to significantly improve prediction accuracy (James et al., 2013). However, if there is a strong feature in the dataset, then the collection of bagged trees will use the strong feature in the first

split, resulting in highly correlated trees. Averaging the predictions from highly correlated trees tends to overfit the data.

Introduced by (Breiman, 2001), random forests decorrelates each tree by only considering a random subset of features m out of p number of features. Hence, some regression trees will not consider the strong feature in construction. Besides m, other hyperparameters of random forests include T, the number of "leaves" in each tree and N, the number of trees in the forest. m, $T$and $N$are selected by cross validation.

5.2.10 Boosted Tree

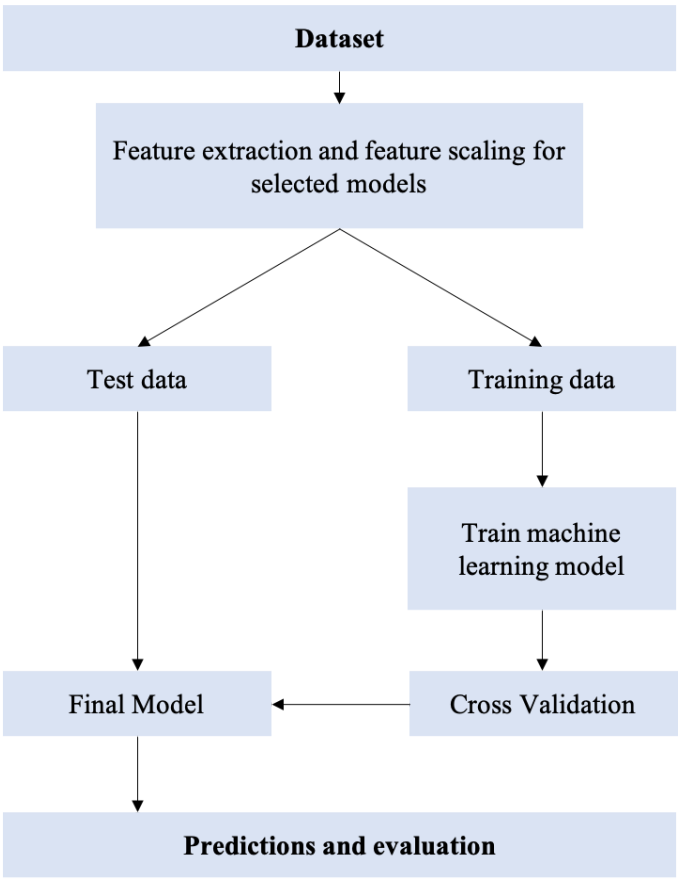
"Boosting" is another approach that improves the predictions from regression trees. "Boosted" trees are grown sequentially using information from previous trees. First, a "shallow" tree is grown from the data (assume d splits = 1). This tree is a weak predictor with high bias because of its simplicity. Next, another tree is grown to fit the prediction residuals of the first tree. The process is repeated until there is a total of N trees in the forest. Forecasts from all the trees in the forest are added together to predict the response variable. By using the information from the residuals of previous trees, the overall model can reduce bias and predict the response variable with increased accuracy.

To reduce overfitting, the learning rate L is specified to control the rate at which the model learns. It has been shown that slow learners can achieve better performance (James et al., 2013). Other hyperparameters include N, the total number of trees and d, the number of splits in each tree. L, $N$and d are selected by cross validation.

5.2.11 Modelling Strategy

Figure 5 presents the modelling strategy for each machine learning model in this paper.

FIGURE 5: MACHINE LEARNING MODELLING FLOWCHART



In the first step, we add squared and interaction features into the dataset, however, due to computational constraints only LASSO (2) and RIDGE (2) models will utilise the additional features. Feature scaling is then performed by standardizing the features for LASSO, RIDGE and ANN models. This is done to ensure that regularization is equally applied across coefficients. The data is then split into the "training" and "test" sets. The "training" data is used to train the machine learning model. 10-fold cross validation is then performed to select the best hyperparameters for the final model. Predictions are then made and evaluated using the "test" data.

EMPIRICAL RESULTS

In this paper, our first aim is to study price determinants in Singapore's public housing market. Our second aim is to evaluate additional insights from the machine learning models. To answer our first aim, we first present the empirical results of the econometric and machine learning models. For the second aim, we examine the feature importances within selected machine learning models.

6.1 Econometric Model

Table 4 presents the summarized empirical results of the OLS model.

Focusing on the main model, we see that the coefficients for variables floor_area_sqm, flat_type, storey_range and remaining_lease are significant and positive. These findings are expected and studied intensively in the literature, hence we will focus on locational variables.

Looking at the transport group of locational features, results indicate that customers value the proximity to MRT and bus stations. By increasing the proximity to the nearest MRT station by 1 kilometre, HDB resale price increases by S$36,843 on average. In section 4.2, it is suggested that shopping centres, government hawkers, bus stations and retail pharmacies are usually located near MRT stations, but their effects on HDB resale prices are not as large, implying that homebuyers place greater value on the convenience of MRT stations rather than other amenities perhaps due to the reliance of MRT as a daily commute to work. In the paper by (Bian et al., 2018), the authors reported a S$150,000 premium for every kilometre increase in proximity. The stark difference in magnitudes could be due to the lack of other amenity controls in their study. For bus stations, resale flat prices increase by S$14,714 on average for every kilometre increase in proximity. The lower magnitude suggests that homebuyers value MRT transportation more than buses, perhaps due to the higher speed of transport and connectivity offered by MRTs. For a kilometre increase in proximity to the nearest bus stop, we see a S$8,230 decrease in resale flat prices on average. As bus stops are usually located on main roads, this suggests that the effect of transport convenience is overshadowed by the level of road noise and congestion in the area.

In the recreation group of locational features, we find that consumers value the proximity to government hawkers, food stalls, shopping centres, supermarkets, community clubs and parks with

various magnitudes. For every kilometre increase in proximity to the nearest government hawker, resale HDB flat prices increase by S$23,135 on average. The convenience of being located near a shopping centre or food stall had a smaller effect, suggesting that hawker centres are the preferred location for food perhaps due to the affordability and variety of food offered there. The findings here contrast the results in the paper by (Bian et al., 2018), who found that the proximity to food courts had a negative effect on real estate prices due to the congestion and noise produced. The only variable that had a negative effect on flat prices was proximity to the nearest child care centre. One explanation could be due to the noise associated with child care centres, affecting the surrounding HDB flats.

In the education group, we find a price premium for an increase in proximity to the nearest primary school and price deficits for the nearest secondary school and junior college. For every kilometre increase in proximity to the nearest primary school, resale HDB flat prices increase by $4,877 on average. For the price deficits caused by secondary schools and junior colleges, one explanation could be due to the fact that these schools have a higher number of students which produced more noise and disturbance. Interestingly, being located within 1 kilometre radius of the top 25 primary and secondary schools had positive price effects of S$2,532 and S$13,999 on average respectively. For primary schools, the positive coefficient is intuitive as children who live within 1 kilometre of the school receive priority allocation, driving up the demand for housing nearby (Ministry of Education Singapore, 2020). Similar results were reported in the paper by (Bian et al., 2018). The coefficient for top secondary schools contrast the findings earlier for general secondary schools in sign and magnitude, this suggests that being located near high ranking schools heavily outweighs the effect of noise in that area. In addition, many top primary schools are affiliated with top secondary schools where the affiliate students are given an admission advantage (Sin, 2017), explaining the higher magnitude compared to top primary schools. For the top 5 junior colleges, the price effect is negative and the magnitude is similar to the findings for general junior colleges. This is expected as junior college admissions are based solely on O-level results and no admission advantages are given.

Lastly, for the health group of amenities, we find surprisingly large price effects of S$19,238 and S$19,946 on average for every kilometre increase in proximity to the nearest clinic and pharmacy respectively, implying that the convenience of being located near medical services is highly valuable to residents. Another interesting finding is the negative effect of being located near a nursing home. For every kilometre increase in proximity to the nearest nursing home, flat prices decrease by S$15,401 on average. This quantifies the negative perceptions buyers have of nursing homes, also known as the

"Not in my backyard" syndrome. To support, there have been instances where residents protested against the building of eldercare facilities in their neighbourhood (Seow, 2017). Given the ageing population in Singapore (Siau, 2017), this finding is significant for policymakers who might have to consider the housing price effects when planning for new nursing home facilities.

Model (1) - (5) serve as robustness checks. Models (1) and (2) check the robustness of physical features. Models (3) - (5) considers a yearly subsample of the data. While there is some evidence of time heterogeneity for bus stops, primary schools, secondary schools and child care centres features, the coefficient signs were unchanged.

6.2 Machine Learning Models

Table 5 in the appendix presents the final configurations of each machine learning model. Table 6A reports the out-of-sample prediction accuracy and R^2values for both OLS and machine learning models. The OLS model reported differs from the econometric analysis as it was estimated using the "training" data instead of the full dataset. The model was then evaluated with the "test" data.

TABLE 6A: OUT-OF-SAMPLE PREDICTION PERFORMANCE BY MODEL

Model	*R squared*	*Adjusted R squared*	*RMSE*	*Gain over OLS*
OLS	0.884	0.884	53.611	-
LASSO (1)	0.883	0.883	53.961	-0.65%
RIDGE (1)	0.885	0.884	53.612	+0.00%
KNN	0.897	0.896	43.671	+18.54%
ANN	0.951	0.951	34.629	+35.41%
RT	0.92	0.919	44.646	+16.72%
Random Forest	0.944	0.943	37.343	+30.34%
Boosted Trees	0.948	0.947	35.860	+33.11%

As seen in Table 6A, both LASSO (1) and RIDGE (1) performed worse in terms of prediction accuracy against the base OLS model. One possible reason could be the strong linear relationship between features and the larger sample size compared to the number of features, as noted by (Melkumova & Shatskikh, 2017).

KNN and Regression Tree (RT) models offer a decent 18.54% and 16.72% improvement in RMSE respectively against the base OLS model. Both models are nonparametric and thus more flexible than OLS as they do not require explicitly making assumptions in the underlying model. However, the trade-off from these models is that they are less interpretable.

ANN performs the best out of all models with a 35.41% improvement over the base OLS model. ANN models are effective in capturing the complex nonlinear dependencies between features (Wang et al., 2016), even without explicitly adding nonlinear features into the model. The remarkable improvement suggests that nonlinear dependencies play a significant role in predicting resale flat prices. However, due to the black-box nature of ANNs, model interpretation is difficult.

Lastly, both Random Forests and Boosted Trees produced prediction scores with a 30.34% and 30.11% improvement over the OLS model respectively. As expected, both ensemble methods performed better than the single Regression Tree by combining many different trees, thus reducing the bias and variance of the model.

TABLE 6B: OUT-OF-SAMPLE PREDICTION PERFORMANCE FOR LASSO AND RIDGE WITH POLYNOMIAL AND INTERACTION TERMS

Model	*R squared*	*Adjusted R squared*	*RMSE*	*Gain over OLS*
OLS	0.884	0.884	53.611	-
LASSO (2)	0.943	0.919	37.741	+29.60%
RIDGE (2)	0.943	0.919	37.788	+29.51%

In Table 6B, we added squared and interaction terms to LASSO (2) and RIDGE (2) models. Prediction performance improved significantly with a 29.60% and 29.51% gain over OLS for LASSO (2) and RIDGE (2) respectively. This proves that nonlinear dependencies between features can indeed improve predictive power. LASSO (2) performed slightly better than RIDGE (2), suggesting that some features were not useful and were removed from the model.

6.3 Comparing Econometric Approach to Machine Learning

It is hard to compare econometric and ML approaches as they are different methodologies tackling different questions, with their own strengths and weaknesses. Econometrics examines the causal relationships in the data and its strength lies in the ease of interpretation and implementation. However, OLS models require explicit assumptions about the variables and the underlying distribution. On the other hand, ML approaches are commonly used to tackle prediction problems. They excel in capturing the patterns in the data to derive accurate predictions but due to its black-box characteristics, interpretation is difficult. However, some machine learning models may complement econometric analysis by examining feature importances, which we will cover in the next section.

6.4 Feature Importances

In Regression Tree, Random Forest and Boosted Tree methods, feature importance is calculated based on the total reduction of RMSE brought by that feature (Guyon & Elisseeff, 2003). The higher the decrease in RMSE, the more important the feature is in improving overall predictive power. This provides additional insight into the determinants of resale flat prices that complement the findings from the hedonic model. Figure 6 shows the feature importances for the 3 ML models. Note that the feature importance within a given model are normalized to sum to one, giving them the interpretation of relative importance. Furthermore, we summed all the relative feature importance across time, planning area and flat model dummies into a single feature respectively.

Looking at the results, all models suggest the floor area to be the most important feature in predicting resale flat prices. For the other physical characteristics, Regression Tree and Random Forests finds that the remaining lease is the second most important physical feature, while Boosted Trees suggest that flat type is more important. All three models agree that the flat model is the least important physical feature. This is expected since flat models are increasingly standardized in newer flats. The planning area is also an important consideration, perhaps due to the varying developmental plans and levels of maturity among planning areas.

FIGURE 6: RANKING OF FEATURE IMPORTANCE BY MODEL

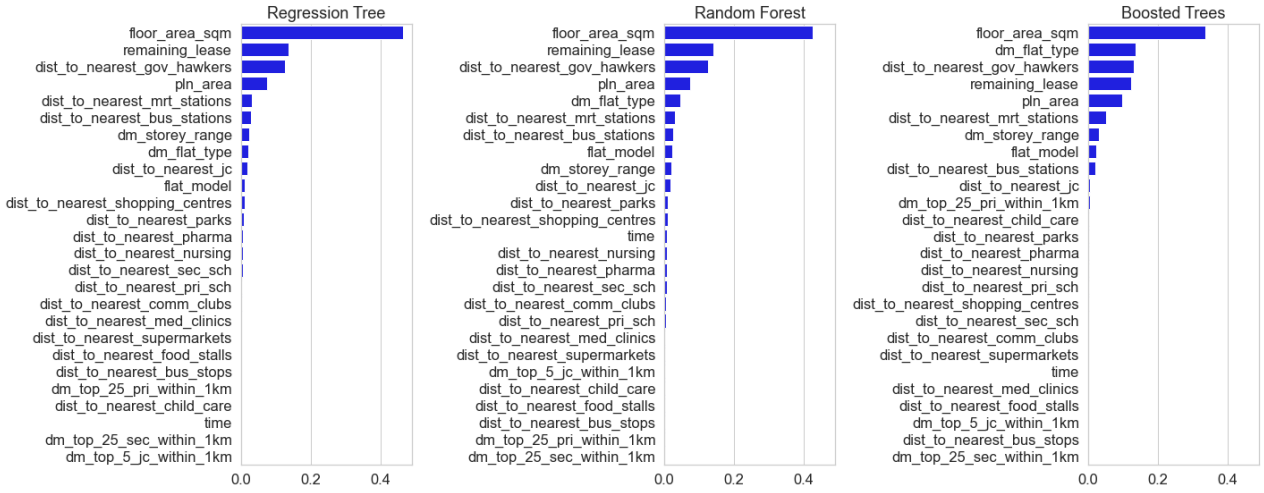
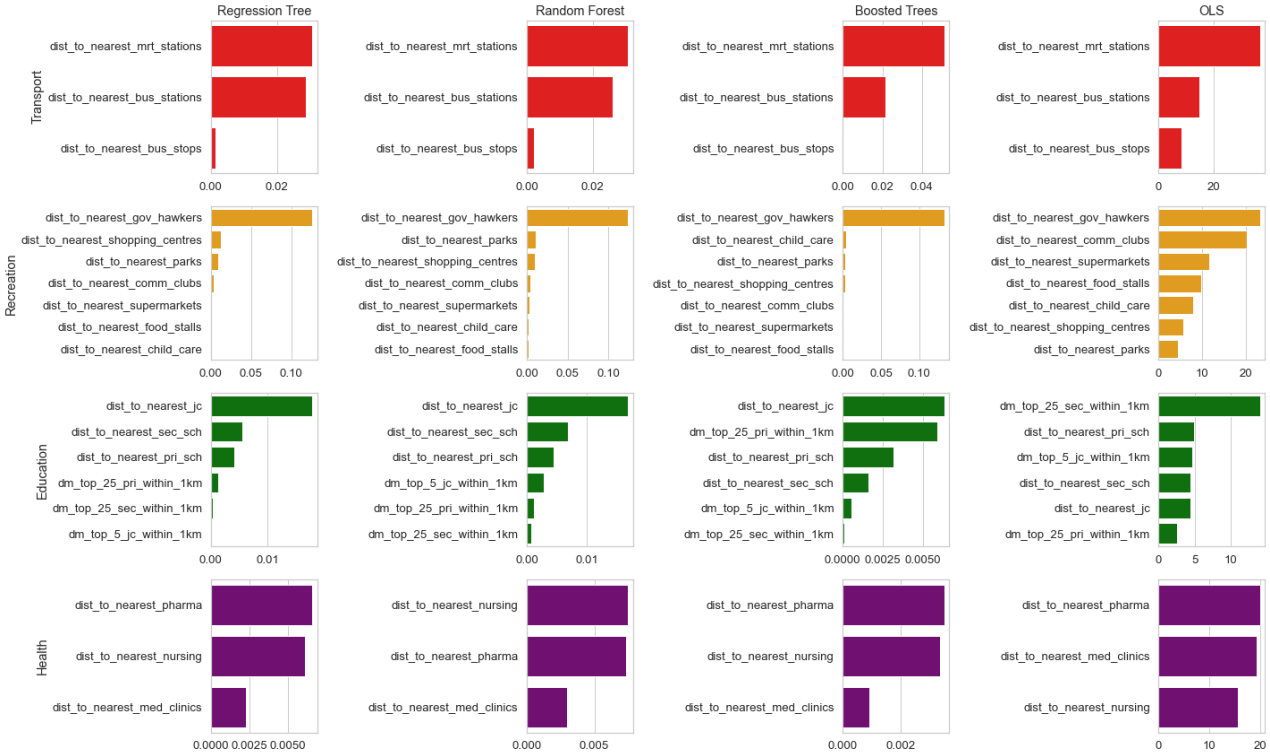


FIGURE 7: RANKING OF FEATURE IMPORTANCES BY AMENITY GROUPS



Next, we examine the locational features in each amenity group in Figure 7. Absolute coefficients from the hedonic model were included for comparison. Row 1 of Figure 7 presents the feature importances for features in the transport group. All 3 ML models find the distance to the nearest MRT station to be the most important among the transport group. Compared to bus stations, the relative importance for bus stops is weaker, implying that the variety of routes (offered in bus stations) is a bigger

consideration for homebuyers. Additionally, the rankings in the group are consistent with the econometric model.

Row 2 of Figure 7 reports the feature importance for the recreation group. Distance to the nearest government hawker is by far the most important feature, suggesting that the accessibility to hawker food was an important consideration for homebuyers. This finding is robust across all models. We also observe that the ML models are not in agreement for the ranking of the other recreational features, perhaps due to the low relative feature importances for the other amenities.

Row 3 of Figure 7 summarises the feature importance within the education group. All 3 ML models find the distance to the nearest junior college as the most important feature in the group. A possible explanation could be that the importances was amplified due to the low number of junior colleges in Singapore. In the econometric model, distance to the nearest secondary school had the largest effect on prices among the educational institutions, contradicting the findings from the ML models. For proximate top schools, both Regression Tree and Boosted Trees suggest that being located near a top ranking primary is more important than a top secondary school or junior college. The result is inconsistent with the econometric model, where top ranking secondary schools had the largest effect on prices.

Lastly, for the health group of locational features as reported in Row 4 of Figure 7, we find that the relative importances for pharmacies and nursing homes are close in magnitude across the ML models. The finding builds evidence suggesting that homebuyers consider the access to medicinal supplies and the proximity to a nursing home as important considerations in purchase.

6.5 Limitations

In this section, we present the limitations of our study. Firstly, the locational data collected was updated by the respective government agencies on varying dates within the period of 2018 - 2020, hence we are unable to ascertain if the locations listed in the data existed throughout the same period, affecting the accuracy of our findings. Secondly, the data considers only transactions between 2018 - 2020 to reduce the impact of the aforementioned problem. We believe that our findings will be more robust with more observations. Lastly, due to computational constraints, we were forced to limit the scope of hyperparameter tuning for each ML model. We were also unable to add polynomial and interaction

terms to all machine learning models, choosing to only add these terms into the LASSO (2) and RIDGE (2) models. Given the computational resources, we believe that more complex models can be trained and tuned to yield better predictive performances.

CONCLUSION

Using transaction data from 2018 - 2020, this paper examines the price determinants of resale HDB flats in Singapore using econometric and machine learning approaches, with a focus on the effects of proximity to 18 locational features across transport, recreation, education and health groups. Our results from the hedonic regression indicate that among all amenities, a kilometre increase in the proximity to the nearest MRT station had the largest premium on flat prices of S$36,843 on average. In the transport group, results show a S$8,230 price decrease for every kilometre increase in proximity to a bus stop, reflecting the negative effect of road noise and congestion in the area. For the recreation group, we find that consumers value access to government hawkers the most, perhaps for the affordable food options and variety. For education, we find that being located near a top 25 primary and secondary school boosted proximate flat prices by S$2,532 and S$13,999 respectively on average. The price premiums could be attributed to the allocation priority for children living within 1 kilometre of the school and the admission advantage given to affiliate students. In the health group, results show that for every kilometre increase in the proximity to the nearest nursing home, flat prices decrease by S$15,601 on average, quantifying the negative perception homebuyers have of nursing homes. We also trained and evaluated 7 machine learning algorithms in this paper. In terms of prediction accuracy, Artificial Neural Networks yielded the best results. We also find that nonlinear dependencies between features could improve predictive power. Looking at the feature importances in the transport group, we find evidence suggesting that the nearest MRT station was an important consideration for homebuyers, consistent with the findings from the econometric model. For recreation, ML and econometric models find that the distance to the nearest government hawker was the most important factor of pricing. For education and health features, findings were less conclusive due to the low relative importances.

Our study is limited by the accuracy of the data and computational constraints as outlined in section 6.5. Nevertheless, we hope that the findings in this paper can provide insight to home buyers and homeowners. We also hope that policymakers would be able to make better decisions in housing policy. Further research can be conducted using updated data and more complex machine learning models to derive greater insights.

BIBLIOGRAPHY

Addae-Duppah, K., & Lan, Y. S. (2010). Shopping Centres and the Price of Proximate Residential Properties. *Pacific Rim Real Estate Society Conference Wellington, New Zealand*. http://www.prres.net/Papers/Addae-Dapaah_Shopping_Centres_Proximate_Residential_Properties.pdf

Ahmed, E., & Mohamed, M. N. (2016). House Price Estimation from Visual and Textual Features. *NCTA, 8th International Conference on Neural Computation Theory and ApplicationsAt: Porto, Portugal*. https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.5220%2F0006040700620068

Athey, S. (2019). The Impact of Machine Learning on Economics. In *THE ECONOMICS OF ARTIFICIAL INTELLIGENCE: AN AGENDA*. University of Chicago Press. https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/impact-machine-learning-economics

Bae, C.-H. C., Jun, M.-J., & Park, H. (2003). The impact of Seoul's subway Line 5 on residential property values. *Transport Policy*, *10*(2), 85-94. https://doi.org/10.1016/S0967-070X(02)00048-3

Baldominos, A., Blanco, I., Moreno, A., Iturrarte, R., Bernardez, O., & Afonso, C. (2018). Identifying Real Estate Opportunities Using Machine Learning. *Applied Sciences*, *8*(2321). doi:10.3390/app8112321

Bengio, Y., & Grandvalet, Y. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *Journal of Machine Learning Research*, *5*, 1089-1105. https://www.jmlr.org/papers/volume5/grandvalet04a/grandvalet04a.pdf

Bian, T., Chen, J., Qu, F., & Jingyi, L. (2018). Comparing Econometric Analyses with Machine Learning Approaches: A Study on Singapore Private Property Market. *Singapore Economic Review*. https://doi.org/10.1142/S0217590820500538

Bowes, D. R., & Ihlanfeldt, K. R. (2001). Identifying the Impacts of Rail Transit Stations on Residential Property Values. *Journal of Urban Economics*, *50*(1), 1-25. https://doi.org/10.1006/juec.2001.2214

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Cao, X., & Hough, J. A. (2012). Hedonic Value of Transit Accessibility: An Empirical Analysis in a Small Urban Area. *Journal of Transportation Research Forum*. DOI: 10.5399/osu/jtrf.47.3.2146

Chiodo, A. J., Hernandex-Murillo, R., & Owyang, M. T. (2010). Nonlinear Effects of School Quality on House Prices. *Federal Reserve Bank of St. Louis Review*, *92*(3), 185-204. https://files.stlouisfed.org/files/htdocs/publications/review/10/05/Chiodo.pdf

Singstat. (2019). Statistics on resident households are compiled by the Singapore Department of Statistics. https://www.singstat.gov.sg/find-data/search-by-theme/households/households/latest-data

Efthymiou, D., & Antoniou, C. (2013). How do transport infrastructure and policies affect house prices and rents? Evidence from Athens, Greece. *Transportation Resarch Part A: Policy and Practice*, *52*, 1-22. https://doi.org/10.1016/j.tra.2013.04.002

Gao, G., Bao, Z., Cao, J., Qin, K., & Sellis, T. (2019). Location-Centered House Price Prediction: A Multi-Task Learning Approach. *arXiv: Learning*. https://arxiv.org/abs/1901.01774

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157-1182. https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

Heyman, A. V., & Sommervoll, D. E. (2019). House prices and relative location. *Cities*, *95*(102373). https://doi.org/10.1016/j.cities.2019.06.004

Ho, W., Tang, B.-S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *28*, 48-70. https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technnometrics*, *12*(1), 55-67. https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634

Honik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366. https://doi.org/10.1016/0893-6080(89)90020-8

Huang, Y. (2019). Predicting Home Value in California, United States via Machine Learning. *STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING*, *7*, 66-74. https://pdfs.semanticscholar.org/5091/5bab05b8c9a1e33ae813b9cd868e78344368.pdf?_ga=2.137497017.802537860.1615402205-1555671675.1615402205

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer. DOI 10.1007/978-1-4614-7138-7

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, *2*, 1137–1143. https://dl.acm.org/doi/10.5555/1643031.1643047

Lancaster, K. J. (1966). A New Approach to Consumer Theory. *Journal of Political Economy*, *74*(2), 132-157. https://www.jstor.org/stable/1828835

Lim, E., & Seah, Y. H. (2017). Win-win scenario in HDB resale market. https://www.businesstimes.com.sg/hub-projects/property-march-2017/win-win-scenario-in-hdb-resale-market

Loukina, A., Zechner, K., Chen, L., & Heilman, M. (2015). Feature selection for automated speech scoring. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, -(-), 12-19. https://www.aclweb.org/anthology/W15-0602.pdf

McCord, J., Mccord, M., McCluskey, W. J., & Davis, P. T. (2014). Effect of public green space on residential property values in Belfast metropolitan area. *Journal of Financial Management of Property and Construction*, *19*(2). https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1108%2FJFMPC-04-2013-0008

Melkumova, L.E., & Shatskikh, S.Y. (2017). Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering*, *201*, 746-755. https://doi.org/10.1016/j.proeng.2017.09.615

Ministry of Education Singapore. (2020, December 21). *How distance affects priority admission*. Ministry of Education Singapore. https://www.moe.gov.sg/primary/p1-registration/distance

Mulley, C., & Tsai, P. (2016). When and how much does new transport infrastructure add to property values? Evidence from the bus rapid transit system in Sydney, Australia. *Transport Policy*, *51*, 15-23. https://doi.org/10.1016/j.tranpol.2016.01.011

Munoz-Raskin, R. (2010). Walking accessibility to bus rapid transit: Does it affect property values? The case of Bogotá, Colombia. *Transport Policy*, *17*(2), 72-84. https://doi.org/10.1016/j.tranpol.2009.11.002

Phang, S.-Y., & Helble, M. (2016). Housing Policies in Singapore. *ADBI Working Paper Series*, (599). https://www.adb.org/sites/default/files/publication/181599/adbi-wp559.pdf

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34-55. https://www.jstor.org/stable/1830899

Seo, Y., & Simons, R. (2009). The Effect of School Quality on Residential Sales Price. *Journal of Real Estate Research*, *31*(3). https://doi.org/10.1080/10835547.2009.12091255

Seow, B. Y. (2017, February 19). A little less Nimby. *The Straits Times*. https://www.straitstimes.com/singapore/housing/a-little-less-nimby

Siau, M. E. (2017, December 6). Elderly to make up almost half of S'pore population by 2050: United Nations Read more at https://www.todayonline.com/singapore/elderly-make-almost-half-spore-population-2050-united-nations. *Today Online*. https://www.todayonline.com/singapore/elderly-make-almost-half-spore-population-2050-united-nations

Sin, Y. (2017, March 7). Parliament: 20% of places in affiliated secondary schools set aside for non-affiliated pupils from 2019. *The Straits Times*. https://www.straitstimes.com/politics/parliament-20-of-places-in-affiliated-secondary-schools-set-aside-for-non-affiliated

Sue, E., & Wong, W.-K. (2010). The political economy of housing prices: Hedonic pricing with regression discontinuity. *Journal of Housing Economics*, *19*, 133-144. http://dx.doi.org/10.1016/j.jhe.2010.04.004

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso (1994). *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, *58*(1), 267-288. JSTOR. http://www.jstor.org/stable/2346178

Wang, L., Chan, F. F., Wang, Y., & Chang, Q. (2016). Predicting public housing prices using delayed neural networks. *2016 IEEE Region 10 Conference (TENCON)*. https://ieeexplore.ieee.org/document/7848726

Wen, H., Jia, S., & Guo, X. (2005). Hedonic price analysis of urban housing: An empirical research on Hangzhou, China. *Journal of Zhejiang University*, *6A*(8), 907-914. https://link.springer.com.remotexs.ntu.edu.sg/content/pdf/10.1631/jzus.2005.A0907.pdf

Wolfgang, B., & Steininger, B. L. (2020). Recent trends in real estate research: a comparison of recent working papers and publications using machine learning algorithms. *Journal of Business Economics*, *90*, 936-974. https://link.springer.com/article/10.1007/s11573-020-01005-w

Wu, J., Wang, M., Li, W., & Peng, J. (2015). Impact of Urban Green Space on Residential Housing Prices: Case Study in Shenzhen. *Journal of Urban Planning and Development*, *141*(4). https://ascelibrary.org/doi/10.1061/%28ASCE%29UP.1943-5444.0000241

Yi, Y., Kim, E., & Choi, E. (2017). Linkage among School Performance, Housing Prices, and Residential Mobility. *Sustainability*, *9*. http://dx.doi.org/10.3390/su9061075

APPENDIX

TABLE 1: SUMMARY OF LOCATIONAL FEATURES

Group	Locational Features	Observations	Source
Transport	MRT Stations	160	Wikipedia
	Bus Stations	25	Wikipedia
	Bus Stops	5,042	LTA DataMall API
Education	Primary Schools	184	Wikipedia
	Secondary Schools	148	Wikipedia
	Junior Colleges	18	Wikipedia
Amenities	Government Hawkers	107	data.gov.sg
	NEA Licensed Food Stalls	36,689	data.gov.sg
	Shopping Centres	173	data.gov.sg
	Supermarkets	478	data.gov.sg
	Community Clubs	119	data.gov.sg
	Parks	350	data.gov.sg
	Child Care Centres	1,920	data.gov.sg
Health	Nursing Homes	74	Healthhub.sg
	Medical Clinics	3,709	Healthhub.sg
	Retail Pharmacy	266	Healthhub.sg

TABLE 2: SUMMARY OF ALL VARIABLES IN THE DATA SET

Variable Name	Definition	Unit
resale_price	HDB resale price	SGD in thousands
floor_area_sqm	HDB floor area	Square Metres
remaining_lease	Remaining lease at time of transaction	Years
dist_to_nearest_mrt_stations	Distance to the nearest MRT station	Kilometres
dist_to_nearest_bus_stations	Distance to the nearest bus station	Kilometres
dist_to_nearest_bus_stops	Distance to the nearest bus stop	Kilometres
dist_to_nearest_pri_sch	Distance to the nearest primary school	Kilometres
dist_to_nearest_sec_sch	Distance to the nearest secondary school	Kilometres
dist_to_nearest_jc	Distance to the nearest junior college	Kilometres
dist_to_nearest_gov_hawkers	Distance to the nearest government hawker centre	Kilometres
dist_to_nearest_food_stalls	Distance to the nearest NEA licensed food stall	Kilometres
dist_to_nearest_shopping_centres	Distance to the nearest shopping centre	Kilometres
dist_to_nearest_supermarkets	Distance to the nearest supermarket	Kilometres
dist_to_nearest_comm_clubs	Distance to the nearest community club	Kilometres
dist_to_nearest_parks	Distance to the nearest park	Kilometres
dist_to_nearest_child_care	Distance to the nearest child care centre	Kilometres
dist_to_nearest_nursing	Distance to the nearest nursing home	Kilometres
dist_to_nearest_med_clinics	Distance to the nearest medical clinic	Kilometres
dist_to_nearest_pharma	Distance to the nearest retail pharmacy	Kilometres
dm_flat_type	Ordinal dummy variable	Integer
dm_storey_range	Ordinal dummy variable	Integer
dm_top_25_pri_within_1km	Dummy variable	Integer
dm_top_25_sec_within_1km	Dummy variable	Integer
dm_top_5_jc_within_1km	Dummy variable	Integer
dm_flat_model	Vector of 19 dummy variables	Integer

dm_pln_area	Vector of 32 dummy variables	Integer
dm_year	Vector of 2 dummy variables	Integer
dm_month	Vector of 11 dummy variables	Integer

TABLE 3: DESCRIPTIVE STATISTICS OF ALL QUANTITATIVE VARIABLES

Variable Name	Count	Mean	STD	Min	Max
resale_price	67092	442.11	155.52	140.00	1258.00
floor_area_sqm	67092	97.71	24.36	31.00	237.00
remaining_lease	67092	74.28	13.25	44.05	97.42
dist_to_nearest_mrt_stations	67092	0.80	0.44	0.02	3.54
dist_to_nearest_bus_stations	67092	1.19	0.65	0.03	4.57
dist_to_nearest_bus_stops	67092	0.12	0.06	0.01	0.38
dist_to_nearest_pri_sch	67092	0.40	0.24	0.04	3.33
dist_to_nearest_sec_sch	67092	0.51	0.29	0.05	3.67
dist_to_nearest_jc	67092	2.68	1.49	0.12	6.90
dist_to_nearest_gov_hawkers	67092	1.92	1.70	0.00	6.12
dist_to_nearest_food_stalls	67092	0.13	0.07	0.00	0.59
dist_to_nearest_shopping_centres	67092	0.68	0.40	0.00	3.21
dist_to_nearest_supermarkets	67092	0.32	0.19	0.00	3.36
dist_to_nearest_comm_clubs	67092	0.48	0.26	0.00	3.91
dist_to_nearest_parks	67092	0.80	0.43	0.04	2.41
dist_to_nearest_child_care	67092	0.13	0.10	0.00	2.96
dist_to_nearest_nursing	67092	0.87	0.43	0.04	2.45
dist_to_nearest_med_clinics	67092	0.18	0.12	0.00	1.12
dist_to_nearest_pharma	67092	0.64	0.34	0.00	2.29

FIGURE 3: MAP OF RESALE HDB FLAT TRANSACTIONS

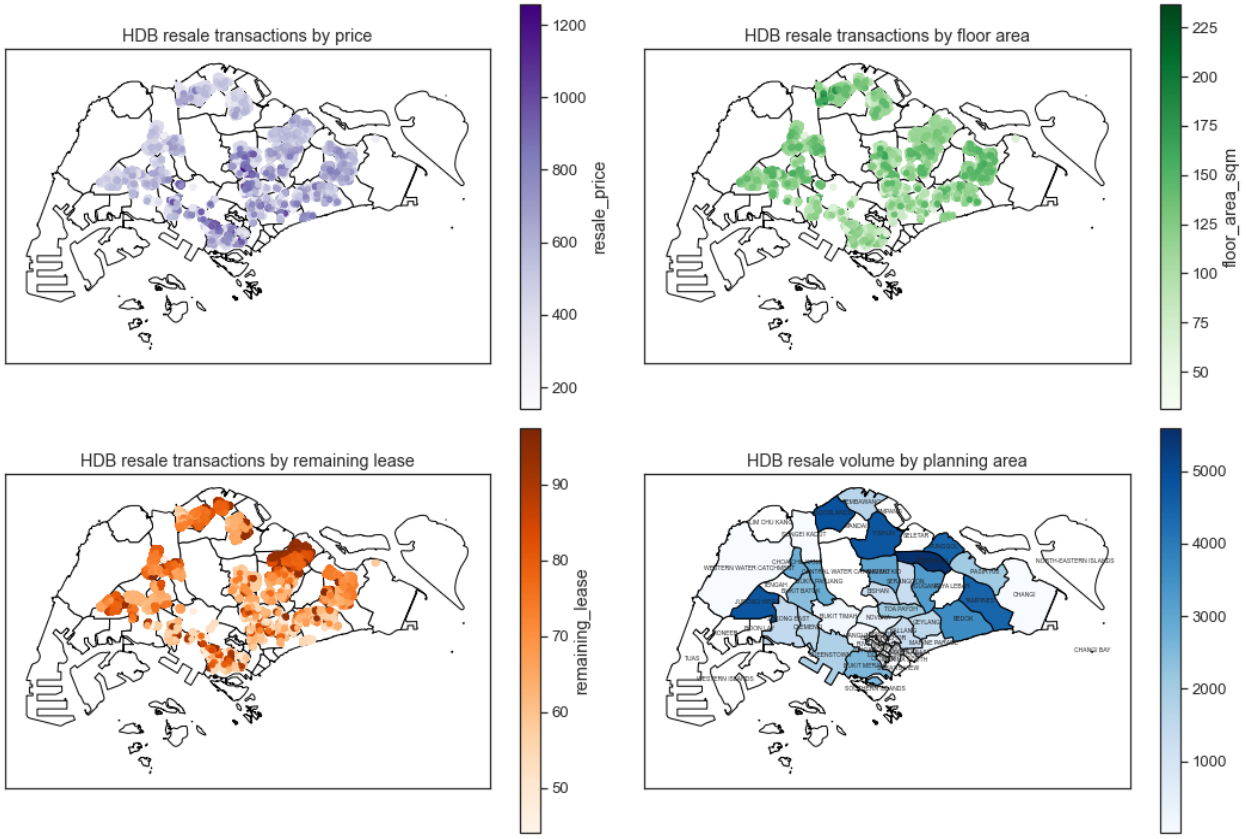


TABLE 4: OLS EMPIRICAL RESULTS

Independent Variable	**Main**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Dependant variable: resale_price						
floor_area_sqm	**3.474*****	3.501***	3.487***	3.483***	3.225***	3.609***
remaining_lease	**5.707*****	4.853***	5.746***	5.399***	5.701***	5.851***
dist_to_nearest_mrt_stations	**-36.843*****	-	-35.996***	-35.230***	-41.592***	-34.750***
dist_to_nearest_bus_stations	**-14.714*****	-	-15.563***	-16.063***	-12.223***	-15.428***
dist_to_nearest_bus_stops	**8.230****	-	8.750**	4.530	2.961	12.918**
dist_to_nearest_pri_sch	**-4.877*****	-	-4.742***	-3.410**	-7.094*	-2.721*
dist_to_nearest_sec_sch	**4.442*****	-	3.836***	4.838***	0.830	5.973***
dist_to_nearest_jc	**4.350*****	-	3.379***	4.841***	4.196***	2.795***
dist_to_nearest_gov_hawkers	**-23.135*****	-	-24.466***	-24.972***	-24.136***	-20.740***
dist_to_nearest_food_stalls	**-9.793*****	-	-7.253**	-15.567***	-10.164*	-5.931
dist_to_nearest_shopping_centres	**-5.570*****	-	-6.246**	-8.265***	-3.283***	-5.565***
dist_to_nearest_supermarkets	**-11.534*****	-	-11.642***	-5.382**	-11.047***	-16.075***
dist_to_nearest_comm_clubs	**-20.134*****	-	-20.351***	-21.977***	-23.476***	-17.973***
dist_to_nearest_parks	**-4.395*****	-	-3.977***	-3.895***	-3.379***	-5.429***
dist_to_nearest_child_care	**7.779*****	-	7.186***	17.135***	1.051	5.092
dist_to_nearest_nursing	**15.601*****	-	15.853***	19.096***	14.300***	12.666***
dist_to_nearest_med_clinics	**-19.228*****	-	-19.384***	-15.178*	-21.834***	-17.297***
dist_to_nearest_pharma	**-19.946*****	-	-18.565***	-22.805***	-18.785***	-17.028***
dm_flat_type	**20.120*****	19.701***	19.697***	18.804***	25.466***	19.729***
dm_storey_range	**56.873*****	56.407***	56.831***	58.190***	57.435***	54.511***
dm_top_25_pri_within_1km	**2.532*****	-	-	2.736**	4.828***	0.477
dm_top_25_sec_within_1km	**13.999*****	-	-	13.469***	14.070***	13.603***
dm_top_5_jc_within_1km	**-4.613*****	-	-	-1.059	-7.709**	-6.764
Timeframe	All	All	All	2018	2019	2020
Number of observations	67,092	67,092	67,092	21,553	22,179	23,360
constant	Yes	Yes	Yes	Yes	Yes	Yes

Controls for flat model, planning area, year and month of transaction	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.889	0.858	0.887	0.893	0.890	0.895
$Adjusted\ R^2$	0.889	0.858	0.887	0.893	0.890	0.894

Note: *** indicates statistical significance at a 1% level of significance. ** indicates statistical significance at a 5% level of significance. * indicates statistical significance at a 10% level of significance. Standard errors used to calculate significance are heteroskedasticity-robust.

TABLE 5: FINAL MODEL CONFIGURATIONS

Machine Learning Model	**Hyperparameters selected by Cross Validation**
LASSO (1)	$\lambda = 0.0991$
LASSO (2)	$\lambda = 0.101$
RIDGE (1)	$\lambda = 68.556$
RIDGE (2)	$\lambda = 5.737$
K-Nearest Neighbour	$K = 5$, Distance = Manhattan Distance
Artificial Neural Network	Number of hidden layers = 1, Hidden layer nodes = (78,), $\lambda = 3$, Activation function = Logistic, Max iterations = 1,000
Regression Tree	$\alpha = 0.29473,\ T = 58,632$
Random Forest	$N = 375,\ m = 87,\ T = 55555$
Boosted Trees	$L = 0.4,\ N = 600,\ d = 3$