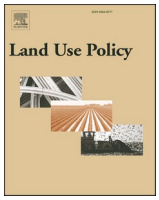
Land Use Policy 123 (2022) 106409

Contents lists available at ScienceDirect

Land Use Policy

journal homepage: www.elsevier.com/locate/landusepol



Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia

Qishuo Gao [*,1], Vivien Shi [2], Christopher Pettit [3], Hoon Han [4]

City Futures Research Center, University of New South Wales, Sydney 2052, Australia

ARTICLE INFO

Keywords:
Property valuation
Machine learning
Hedonic price model
Sub-area

ABSTRACT

Property valuation plays a significant role in urban economics and is of great importance to various stakeholders who interact and shape the city, including property owners, buyers, banks, land developers, real estate agents, local councils and government planning authorities. In the literature, various predictive models have been proposed to automate the calculation of property value, most of which endeavour to factor in the combination of property characteristics, market factors and location-based attributes associated with individual properties use large citywide databases. At the same time, it has been widely acknowledged that regional sub-areas have impacts on property price prediction. Therefore, this paper aims to investigate the performance of various techniques on sub-areas using the Greater Sydney Region as the study area. The sub-area in this paper is defined as the statistical areas (SAs) as defined by the Australian Bureau of Statistics. In particular, two different SA geographies (SA4, SA3) along with the City Level are adopted to understand the spatial dependence which occurs at different levels. With real-world transaction records and data collected from a diverse range of sources, various methods including the traditional hedonic price model (HPM) and popular machine learning (ML) approaches are implemented and evaluated for property price prediction. Two different property markets for residential property are modelled, being for housing stock and apartment (unit) stock. Experimental results show that Random Forest and Gradient Boosting-based methods outperform other approaches in most scenarios and that the high spatial resolution property sub-area (SA3) improved the performance in terms of overall model accuracy. This research provides insights into how sub-area machine learning models can be employed in real estate to characterize property price, and helps understand the influential factors in different local geographical areas for policy-making.

1. Introduction

Property markets are a key component to the fundamental economics underpinning cities where housing affordability, accessibility and affluence are key indicators to a city's performance. The residential property market comprises a significant part of the overall property market, and an accurate property valuation plays an important role in financial activities as well as policy decision in urban planning. For example, banks use valuation to calculate the loan to value ratio to decide the loan amount to be approved. Land developers consider the market price of properties in determining the revenue in order to calculate the return on investment and planning agencies are interested in property valuations in order to calculate the uplift of value and determine land developer contributions and betterment (Huston and Lahbash, 2018; Lee and Locke, 2020). Nowadays, although research into real estate has developed significantly, the price estimation in the housing market still significantly relies on on-site inspection and manual review from professional valuers. Real assets have more complex dependencies in terms of valuation, such as their inherent attributes, market volatility, and neighbourhood characteristics (Schulz and Werwatz, 2004; Soltani et al., 2021). The nature of spatio-temporal factors also makes an accurate prediction very challenging. Over the past

* Corresponding author.
E-mail addresses: qishuo.gao@unsw.edu.au (Q. Gao), ye.shi@unsw.edu.au (V. Shi), c.pettit@unsw.edu.au (C. Pettit), h.han@unsw.edu.au (H. Han).
[1] 0000-0002-9249-4065.
[2] 0000-0002-9100-2751.
[3] 0000-0002-1328-9830.
[4] 0000-0003-3200-7728.

https://doi.org/10.1016/j.landusepol.2022.106409
Received 17 July 2021; Received in revised form 8 August 2022; Accepted 20 October 2022
Available online 25 October 2022

decade, advances in urban imagery products and open data government releases have led to an increasing amount of fine-scale city data. This has enabled researchers to developed more sophisticated predictive property price models and scenario planning tools (Pettit et al., 2020). However, leveraging the advances in computational technology and the advent of "big data" is still an ongoing issue and opportunity (Bourassa et al., 2020). This research aims to study residential property valuation by taking advantage of various sources of data and novel modelling methods afforded through machine learning (ML) techniques (hereafter the term 'property valuation' will be used instead of 'residential property valuation').

The estimation of property value is determined by a bundle of property characteristics, locational features and market factors (Soltani et al., 2021). The property characteristics are mainly the dwelling attributes, such as the number of bedrooms, property age, land size, and building type etc. The locational features reflect the neighbourhood environment as well as accessibility factors (Azmoodeh et al., 2020; Wen et al., 2018; Wu et al., 2019). For example, the distance to the central business district (CBD) is usually inversely related to the urban housing price (Filippova and Sheng, 2020). The housing location is also related to the accessibility to public services and the attraction to amenities (Dai et al., 2016; Diaz and Mclean, 1999; Evangelio et al., 2019; Li et al., 2019; Pagliara and Papa, 2011; Song et al., 2019; Xu et al., 2016; Yang et al., 2019), such as public transportation, schools, hospitals, and shopping centres. The environmental factors, such as walkability (Kim and Kim, 2020), crime rate (Buonanno et al., 2013; Case and Mayer, 1996; Ihlanfeldt and Mayock, 2010; Thaler, 1978; Tita et al., 2006), green space (Liebelt et al., 2019; Mansfield et al., 2005; Noor et al., 2015; Ye et al., 2019), zoning (Bento et al., 2009; Glaeser and Gyourko, 2002; Kendall and Tulip, 2018; Maser et al., 1977), environmental risk (e.g. flood and bushfire) (Bin and Landry, 2013; Giglio et al., 2015; Rojas et al., 2013), have been considered in studying the housing market. In addition, it has been acknowledged that market factors (Englund and Ioannides, 1997; Farlow, 2005; Rahman and Masih, 2014; Xu, 2017) including Gross Domestic Product, (GDP), interest rates, property price index, and inflation rate impact the property price at a macroeconomic level. The choice of variables to be included is highly dependent on the availability of specific data, as well as the accuracy and the granularity of these data sources.

Hedonic price models (HPM) (Chung-Ang, 2019; Lancaster, 1966; Rosen, 1974) have been widely used in property appraisal and quantifying the impacts of various factors on property prices, usually in a form of multiple regression where the sale price is considered as the dependent variable and the hedonic attributes as explanatory variables. Standard Ordinary Least Square (OLS) (Rosen, 1974) regression in a semi-log functional format has been the most frequently used statistical technique in past studies (Pettit et al., 2020). OLS is simple to understand and implement, and the visualization of the coefficients of each variable makes the model easily interpretable and the results transparent. However, OLS is sensitive to outliers and the performance would be compromised if the independent variables are correlated (Anselin, 2013; Chica-Olmo et al., 2019).

In the era of big data, ML has received a lot of attention as a means of large-scale data analysis (Alpaydin, 2020). Experts in the field of economics have been using ML algorithms to estimate property values extensively over the past decade (Fan et al., 2018; Mohd et al., 2020; Mu et al., 2014; Mullainathan and Spiess, 2017; Zulkifley et al., 2020). The advantages of these algorithms are their ability to output predictions from large datasets with more precision than traditional regression models. Support Vector Machine (SVM) (Wang et al., 2014), Decision Trees (Fan et al., 2006), Random Forest (RF) (Antipov and Pokryshevskaya, 2012; Čeh et al., 2018; Hong et al., 2020), Gradient Boosting Models (GBM) (Jain et al., 2019; Mohd et al., 2020; Peng et al., 2019), and Neural Networks (Limsombunchai, 2004) are the most frequently applied ML models in property valuation and could provide accurate prediction results. Convolutional neural networks (CNNs) have the privilege of extracting features from imagery and thus able to collect more information for property valuation (Ge, 2019; Law et al., 2019; Piao et al., 2019). Also, recurrent neural networks (RNNs) (Chen et al., 2017) have been used to better capture the spatio-temporal patterns in housing market while providing the accurate prediction on property prices. The above-mentioned approaches have also been combined in the automated valuation process (Ge, 2019; Lu et al., 2017; Zhao et al., 2019).

Most of the existing modelling endeavours calculate the property valuation with the assumption that the study area is one homogenous market, rather than a series of housing submarkets which in reality is more aligned to how cities function economically (Randolph and Tice, 2013). The traditional global approach assumes that the weights of different attributes are constant in the modelling process across the whole study area. It neglects the spatial heterogeneity of the contributions of the factors in local areas and different housing market segmentations. It was argued that spatial dependence exists across geographical areas where nearby properties may share similar characteristics, such as the structural features, the attraction to amenities and the locational public service, hence the underlying relationships between these features and housing market might vary across space (Bourassa et al., 2010; Se Can and Megbolugbe, 1997). For example, green space may have greater impact on house price in urban areas than suburban areas, and education facilities would account much more for house price in school districts. Transportation may play different roles in regional areas and inner cities regarding house prices. Therefore, housing analysis implemented on a broader geographical scale may provide suboptimal results as the local dynamics are often ignored in the process (Worthington and Higgs, 2013).

Over the last a few decades, hedonic calibration models have been developed, such as spatial autoregressive models (e.g., spatial error and spatial lag models), to control for spatial dependence (Armstrong and Rodriguez, 2006; Haider and Miller, 2000; Löchl and Axhausen, 2010). Also, geographically weighted regression (GWR) has been proposed to explore spatial non-stationarity between property values and predictors (Du and Mulley, 2006, 2012; Dziauddin et al., 2015; Mulley et al., 2016). In addition to the local effects in space, temporal dimension (i.e. dependence in time) has also been considered in spatial analysis, for example, geographically weighted Temporal Regression (GTWR) model accounts for varying local spatial effects across time based on a spatio-temporal weight matrix between observations (Fotheringham et al., 2015). Some other methods have been also applied to deal with spatial dependence. For instance, Se Can and Megbolugbe (1997) constructed a house price index measuring the spatial dependence via the structural attributes, and then the house price index was included in their hedonic models. The authors in (Fik et al., 2003) presented an interactive variable approach and tested its ability in explaining price variations. The approach was constructed based on the given location and the similarities of locational characteristics. Thibodeau (2003) examined several approaches to delineate within-metropolitan-area housing submarket boundaries and employed spatial econometric techniques to spatially adjust the OLS method for house price prediction.

There have also been other ways to take into account the spatial heterogeneity issue, such as splitting the study area into smaller geographic scales and developing models in each sub-area in parallel. Calibrating models in sub-areas could directly capture the effects of house characteristics in local areas. For example, Neill et al. (2007) included sub-area (census tract) variables in OLS and also implemented geostatistical methods to study the impacts of air quality variations on house price in Las Vegas. Case et al. (2004) divided a geographic housing market into districts and developed property valuation models separately to take advantage of the neighbourhood effects on house price. The results in both research show that dividing a global geographical area into small areas performs better than traditional OLSs. On another aspect, dividing the large sample into smaller units is significantly more computationally efficient for large-scale analysis, especially for those

algorithms which rely on intensive calculation, such as GWR (Feuillet et al., 2018). It is important to note that ML algorithms identify the housing price patterns from a given set of variables and are able to model non-linear effects, thereby resulting in more accurate predictions (Mohri et al., 2018). However, the spatial dependencies are still ignored if the models are performed on the whole geographical area. There is limited research that has examined the ML capability in the context of developing models across sub-areas. This is one of the research gaps this paper endeavours to address.

In this research, we explore the modelling of automated property valuations in Greater Sydney, Australia, by applying a variety of methods at different geographic scales as defined as the Australian Bureau of Statistics (ABS) statistical area (SA) levels.[5] Statistical Area (SAs) are based on the concept of a functional area within which people commute and conduct daily activities and the classification of the SA areas is based on population, economy, and infrastructure. ABS collects, releases and analyses census data within the SA areas. The SAs are often taken as the geographic segmentation of housing markets for generating house price indices in the industry (CoreLogic, 2018).

Previous studies have considered the SA as an indicator of housing submarket in Australia and used them in property valuation models (Gao et al., 2019). In this study, we compared the performance of different ML techniques in property valuation models as applied in Sydney at different geographic scale (SA3, SA4 and Greater Sydney Metropolitan Region) in order to capture the neighbourhood characterizes and spatial dependence more efficiently, especially for properties that have similar proximity and neighbourhood features. Various methods are (11 in total) applied to perform the modelling, including Linear regressions (i.e. OLS, Ridge, Lasso), support vector regression (SVR), tree-based algorithms (i.e. Decision Trees, Random Forest), Gradient Boosting-based methodologies (i.e. Gradient Boosting Model, XGBoost), and neural network-based method (i.e. Multilayer Perceptron). In addition, a local form of linear regression – GWR is run across Sydney and calibrated with an adaptive bandwidth. Finally, the performance of all the models is validated and compared using a clearly defined set of accuracy metrics that are used either in research or industry.

In summary, this research provides a number of methodological contributions to the field of property valuation modelling. Importantly, the property types of House and Apartment (Unit) are both modelled separately in this research. Most of the existing literature focuses on only house price prediction, while Apartment is nowadays occupying a large portion of the real estate market in the world. In Australia, Apartments have become an important part of the housing mix. The number of Apartments constructed has tripled each year since 2009 (Rosewall and Shoory, 2017), and the increase in Apartment construction and the location of this Apartment stock across Cities in Australia, particularly Sydney is resulting in high density living and urban agglomeration. Hence the separation of Housing and Apartment models for calculating property valuation is important. In this study, we examine the capability of a suite of property valuation models in predicting both Houses and Apartment prices and compare the performance of these models for these two distinct property markets. Another distinctive contribution of this research is the implementation and evaluation of property price models from a sub-area perspective. In addition, we analyze the importance of the selected variables in influencing the property price and illustrate that the variables' contributions to the property price vary for different geographical areas. Finally, this study incorporates both in-time and out-of-time strategies for evaluating property valuation model performance.

The remainder of this paper is organized as follows. Section 2 gives a brief review of the methodology used in this research. The study area and data description are outlined in Section 3. The experimental design, results, and findings are described in Section 4 and Section 5, respectively. Section 6 provides concluding remarks and outlines proposed future work.

2. Methodology

In this paper, several property price prediction models have been developed and evaluated using a suite of modelling approaches including Linear regression, support vector regression, tree-based methods, Gradient Boosting-based approaches, and sequential neural network.

2.1. Linear regression model

Hedonic price model (HPM) has been widely used in predicting property price. HPM has different functional forms, such as linear, semi-log, and double-log (Lancaster, 1966; Rosen, 1974). As a basic regression problem, linear regression can be represented as:

$$y_i = \beta_0 + \sum_{k=1} \beta_k x_{ik} + \varepsilon_i \quad i \in 1, 2, ., n \tag{1}$$

where y_i is the sale price of i sample, x_{ik} denotes the value of k feature for i sample; β_0 represents the model constant, β_k with $k = 1, 2, ., K$ is the weight coefficient for i sample, and ε_i refers to the effort component reflecting the difference between an individual observed response and prediction from the model due to random factors, n is the total count of samples.

The ordinary least square (OLS) regression has been the most widely applied techniques for hedonic models, and the semi-log transformation is routinely used as it could be used for skewed data and its coefficients could be easily interpreted and also (Lancaster, 1966; Rosen, 1974).

Least absolute shrinkage and selection operator (Lasso) has been considered as a standard linear regression algorithm and it performs the regression with $L1$ regularisation term. Lasso algorithm can also be used in a variable selection manner as the $L1$ penalty forces certain coefficients to be set as zero so that some of the variables would not contribute to the predictions. In this way, important features have large coefficients. However, the performance of Lasso highly depends on the selection of the regularisation parameter, which would lead to low performance if set as a large value, and a small value may cause overfitting (Tibshirani, 1996).

Ridge regression (McDonald, 2009) enforces the coefficients to be small if the features have a small influence instead of assigning zero values, thus keeps all variables in the model. Ridge works well when the data suffers from multicollinearity, but it would introduce bias to the parameter estimates if there is a large set of variables (McDonald, 2009).

Another linear algorithm applied in this paper is Elastic Net regression (Zou and Hastie, 2005), which uses both $L1$ and $L2$ regularisation terms. The fine-tuning of both regularisation terms increases the computational cost, and the flexibility of choosing two parameters is likely to introduce overfitting.

The local spatial regression motel, GWR, is also included in this paper to compare with other models run at sub area level. GWR was developed by (Du and Mulley, 2006; Fotheringham et al., 2003) firstly applied it in the public transportation infrastructure to capture the value uplift in the Northeast of England for light rail. It allows local global parameters to be estimated by embodying spatial coordinates into the traditional global regression model, so that the model is rewritten as:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \tag{2}$$

where (u_i, v_i) denotes the coordinates of the ith point in space and $\beta_k(u_i, v_i)$ is a realization of the continuous function $\beta_k(u, v)$ at point i.

[5] https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Digital+Boundaries.

Weighted least squares provide a basis for understanding how GWR operates. In GWR an observation is weighted in accordance with its proximity to location i so that the weighting of an observation is no longer constant in the calibration but varies with. Data from observations close to i are weighted more than data from observations farther away, that is,

$$\check{\beta} \quad (u_i, v_i) = \left(X^T W(u_i, v_i) X\right)^{-1} X^T W(u_i, v_i) y \tag{3}$$

where $\check{\beta}$ represents an estimate of β, and $W(u_i, v_i)$ is an n by n matrix whose off-diagonal elements are zero and whose diagonal elements denote the geographical weighting of each of the n observed data for regression point i. The global regression model is a special case of GWR in which the parameters are spatially invariant (Fotheringham et al., 2000).

2.2. Support vector regression

Support Vector Regression (SVR) (Awad and Khanna, 2015) uses the same principle of support vector machine (SVM), and adapts the classification into a regression problem. For samples that are not linearly separable, kernels can be applied to transform the data into a higher feature space to make it possible to perform the linear separation. There are some kernels can be applied for this purpose, such as polynomial kernels, intersection kernels and string kernels. Gaussian kernel is the most popular kernel by far. Some software packages, such as Liblinear (Fan et al., 2008) and LibSVM (Chang and Lin, 2011), have been commonly used to perform SVM and optimize for parameters.

2.3. Tree-based model

Tree-based algorithms are considered one of the most widely used supervised machine learning methods. These algorithms empower the models with higher prediction accuracy, robustness, and ease of interpretation. Two tree-based methods were applied in this paper: Decision Tree and Random Forest.

A decision tree (Myles et al., 2004) uses a tree-like graph to highlight all possible results of a decision. It is a support tool that an algorithm is displayed as conditional control statements. Fig. 1 shows a simple example for the decision tree. Nodes represent conditions, and the model goes left or right on the basis of the condition. For regression problem, a decision tree trains the model by observing the features of data in a tree structure and produces a continuous output.

Decision trees are sensitive to data and the prediction results would be quite different with different training samples, which would compromise the robustness of a model. Random forest has been widely used to replace decision trees in both classification and regression problems, as it shows the power of combining multiple decision trees into one model.

Random forest (Biau and Scornet, 2016) aggregates many decision trees and there is no interaction between those trees during training. The output of a random forest is the average of the predictions from each tree. Fig. 2 illustrates the principle of a random forest, and each tree draws a random feature from the data set for splitting which avoids the overfitting of the algorithm.

2.4. Gradient boosting-based model

In this study, gradient boosting-based approaches were also applied in the property valuation. The main idea of a boosting model is to convert a weak learner into a strong one, and usually a decision tree is used as the weak learner. Gradient Boosting trains many tree algorithms in an additive and sequential manner. Fig. 3 is an example of a gradient boosting algorithm, and each new tree is a fit on the original data set.

During each iteration of training, the outputs of the weaker are compared to the expected values, and a gradient method is used to minimize the loss function.

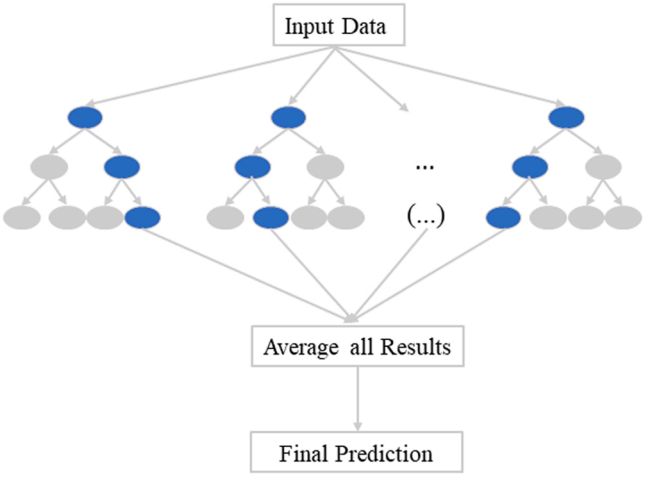


Fig. 2. The structure of a Random Forest.

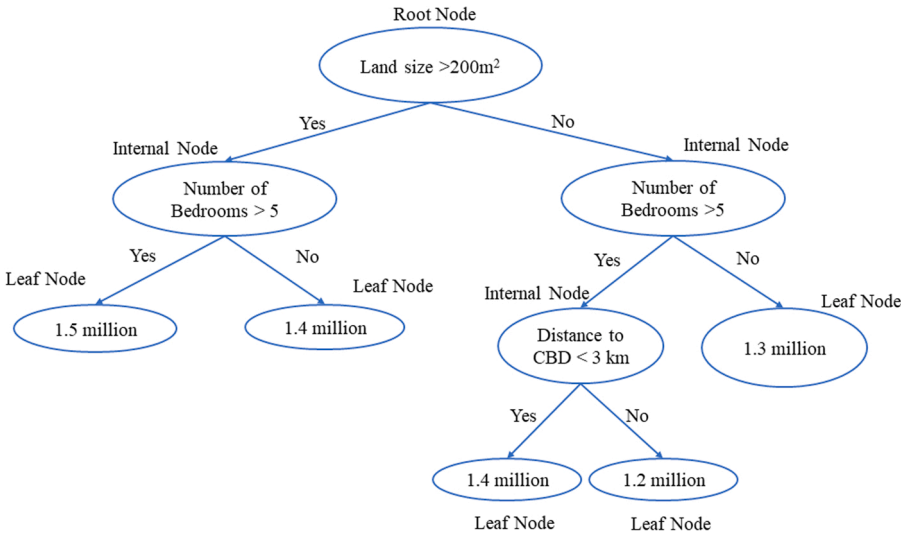


Fig. 1. The illustration of a decision tree.

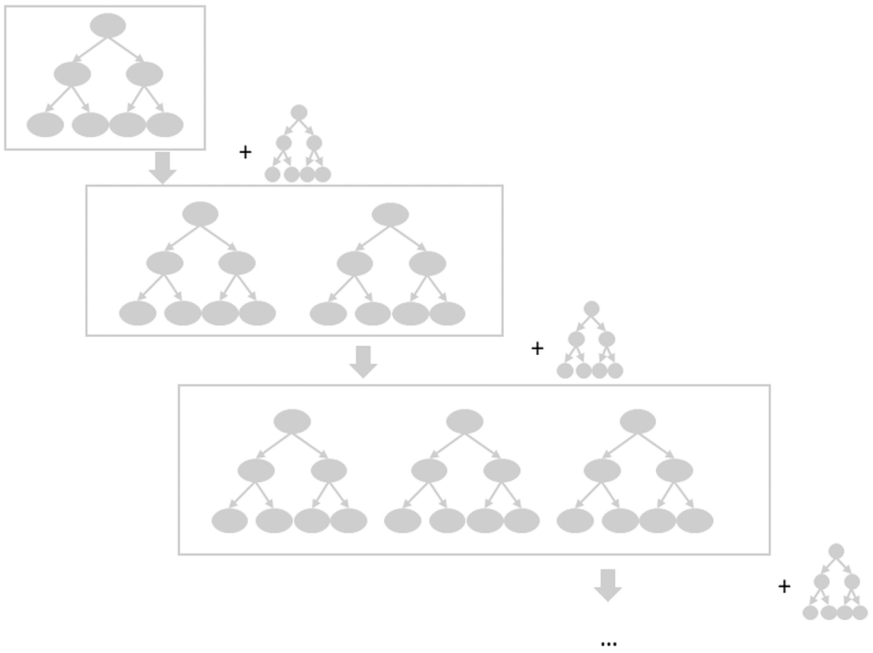


Fig. 3. The architecture of a gradient boosting model.

XGBoost (Chen and Guestrin, 2016; Chen et al., 2015) stands for Extreme Gradient Boosting which is a specific implementation of the gradient boosting method. Compared to traditional gradient boosting, XGBoost applies more approximations for a better model built. XGBoost computes the second-order gradients and applies more advanced regularisation which improves the model generalization. However, the application of each model highly depends on the specific data set. For comparison, both the classic gradient boosting model (GBM) and XGBoost were tested in this study.

2.5. Neural networks

Deep learning has attracted a lot of attention in the field of ML in recent years due to its superior performance compared to traditional statistical approaches. The basic idea behind deep learning is to utilize a hierarchical level of artificial neural networks (ANN) (Anderson, 1995), which is usually called neural networks, for automatically learning features via layers. From Fig. 4, it can be seen that a basic neural network maps the input (e.g housing features) to output (e.g. housing price) via hidden layers, and a hidden layer consists of several neurons.

Deep learning networks are more distinguished from few layers network by the "depth" of the network. A deep network can have as many as layers. Additionally, an active function, such as sigmoid function, rectified linear unit (ReLU), and softmax, is usually used after layers to perform the nonlinear transformation. The network shown in Fig. 4 is a *feedforward* neural network as the output of one layer is subsequently used as input to the next layer. The class of such neural networks are also referred to as multilayer perceptron (MLP), which could also be applied in property valuation and has been included in this study.

3. Study area and data description

In undertaking this study, we collected data from various sources across the Greater Sydney Metropolitan area, including the property transaction data with property structure features, locational characteristics, and census.

3.1. Historical sales and property characteristics

The property sales data was provided by Australian Property Monitors[2], one of the largest online property sales advertising portals in Australia. Property sales in 2018 were extracted for the Greater Sydney Metropolitan area. There were in total of 77, 063 sales being geocoded from the original data.Separate models are developed based on the market segmentation by property type: House and Apartment (Unit). According to the property classification code, there are 54, 084 sales records classified as House, and 22, 979 sales records are categorized as Unit.

The sales data contains several property-related variables that enable

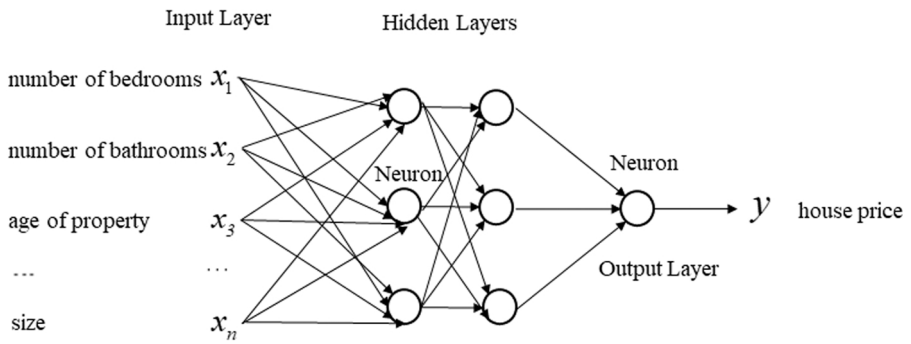


Fig. 4. An illustration of a simple neural network.

extensive modelling of the housing characteristics. The main variables in the transaction records are the transaction time, sold price, land area size (only available for House), dwelling type (only available for House, and it has been categorized into House, Semi-detached, Townhouse, and Villa), number of bedrooms, number of bathrooms, number of garages, and geographical coordinates. Attributes with large portion of missing value have been removed, such as swimming pool, solar panel, air-conditioning, and heating.

3.2. Locational information

The case study properties were geocoded using their coordinate information allowing for measuring distances to different destinations. Several locational variables were also considered for modelling, and most of them are the point of interest (POI) data. The POI files originally received in a shapefile format, and then point-to-point distances are measured. The distances to each POI are calculated at property level and treated as additional variables for modelling. Table 1 lists the derived variables and data sources.

3.3. Census and statistical data

Table 2 illustrates the ABS census data on neighbourhood socioeconomic characteristics used in the predictive property models.

4. Experimental design

In this study, all the data processing and models were implemented in Python 3.8. Several powerful packages have been installed to perform the techniques, such as Scikit-learn, Tensorflow and XGBoost. Scikit-learn (Pedregosa et al., 2011) features various supervised and unsupervised algorithms such as SVM, random forest, and linear regressions. Tensorflow (Abadi et al., 2016) originally developed by the Google Brain team, has been designed as an end-to-end platform for ML algorithms, especially neural networks. XGBoost (Chen and Guestrin, 2016) is an efficient and flexible library for implementing different gradient boosting frameworks. The computing specifications used to support the predictive property modelling reported in this paper is Intel (R) Core (TM) i7–8665U CPU 1.90 GHz and 32 GB Installed Memory.

Table 1
POI data and their data sources.

Data Source	Data Derived	Variable Name	Variable Defination
NSW Land & Property Information (LPI)	Railway stations Shopping centers Hospitals Universities Beach Coast Community Swimming pool City Centers Transmission lines Main road	Distance to Railway Station Distance to Shopping Center Distance to Hospital Distance to University Distance to Beach Distance to Coast Distance to Community Swimming Pool Distance to City Center Distance to Transmission Line Distance to Main Road	The distance calculated from the property to the nearset railway station The distance calculated from the property to the nearset shopping center The distance calculated from the property to the nearset hospital The distance calculated from the property to the nearset University The distance calculated from the property to the nearset Beach The distance calculated from the property to the nearset Coast The distance calculated from the property to the Community The distance calculated from the property to the nearset Swimming Pool The distance calculated from the property to the nearset City Center The distance calculated from the property to the Transmission Line The distance calculated from the property to the Main Road
NSW Department of Education (DET)	Primary school location High school location	Distance to Primary School Distance to High School	The distance calculated from the property to the nearset Primary School The distance calculated from the property to the nearset High School

Table 2
Data sources and the names of the census and statistical variables.

Data Source	Data Derived	Variable Name	Variable Defination
Australia Bureau of Statistic (ABS)	Digital Statistical Area, boundaries and demographics, including population and age, occupation, household income, country of birth.	Percentage of Population Born Overseas Median Weekly Family Income Percentage of Population Aged 65 above Percentage of people with high income Percentage of people unemployed	Percentage of people living in Australia but born overseas Median weekly gross household income Percentage of residents aged 65 and above Percentage of people earn AUD 2000 and more weekly Percentage of people unemployed as either full-time or part-time
NSW Bureau of Crime Statistics and Research (BOCSAR)	Recorded crime dataset	Crime Rate	Number of crimes/ Population

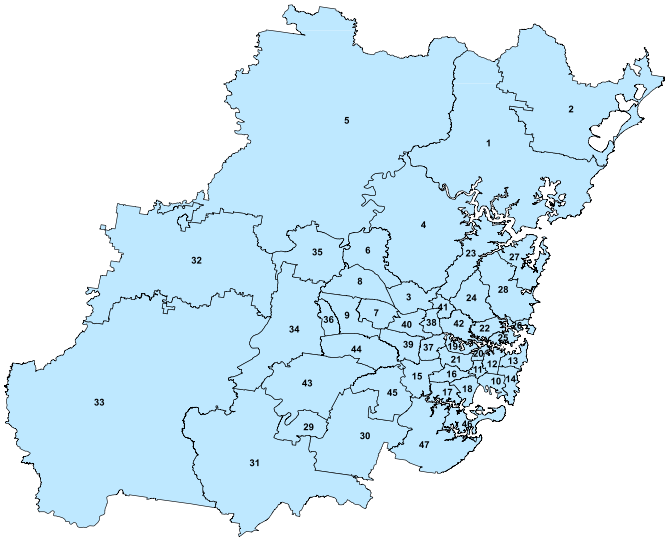
*The data derived in Table 2 are the smallest spatial aggregation level possible for each property, for example, postcode level/SA1 level and will derived from upper level if there is no data available.

4.1. Statistical area levels

Models were developed at different SA levels: Statistical Area Level 3 (SA3), Statistical Area Level 4 (SA4), and also the Greater Sydney (GSYD) City Level. The SA levels applied are followed by the main structure of the Australian Statistical Geography Standard, which is comprised of seven hierarchical levels: Mesh Block (MB), Statistical Area Level 1 (SA1), Statistical Area Level 2 (SA2), SA3, SA4, State and Territory (S/T), and Australia (AUS). Each level directly aggregates to the level above. There are in total 47 SA3 and 15 SA4 areas in the GSYD Metropolitan area.

In this study, models were calibrated on each SA3, SA4, and GSYD Metropolitan region in parallel. SA1 and SA2 areas were skipped for analysis to keep a reasonable granularity level and allow for sufficient training sample in each local area. In addition, an adaptive multi-level framework is also developed for sample selection: if there are not sufficient training samples (<100) in an SA3 area, the model will be carried out in the corresponding SA4 area, and the model will be developed across GSYD if the training samples are less than 100 in the respective SA4 area. The threshold is set as 100based on the sensitive test which is shown in Section 5.3. The criteria were applied to ensure the sub-area models were developed based on a valid and reasonable sample size. Figs. 5 and 6 illustrate the digital boundaries and the distribution of training samples of House and Apartments (Unit) at the SA3 level and SA4 level, respectively.

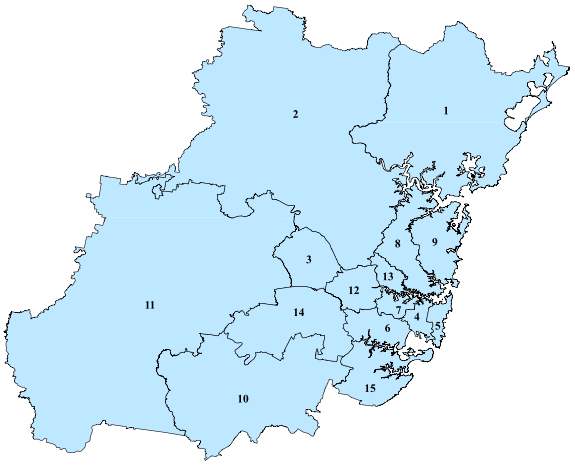
(a)



SA3 Name

1 Gosford	17 Hurstville	33 Blue Mountains - South
2 Wyong	18 Kogarah - Rockdale	34 Penrith
3 Baulkham Hills	19 Canada Bay	35 Richmond - Windsor
4 Dural - Wisemans Ferry	20 Leichhardt	36 St Marys
5 Hawkesbury	21 Strathfield - Burwood - Ashfield	37 Auburn
6 Rouse Hill - McGraths Hill	22 Chatswood - Lane Cove	38 Carlingford
7 Blacktown	23 Hornsby	39 Merrylands - Guildford
8 Blacktown - North	24 Ku-ring-gai	40 Parramatta
9 Mount Druitt	25 North Sydney - Mosman	41 Pennant Hills - Epping
10 Botany	26 Manly	42 Ryde - Hunters Hill
11 Marrickville - Sydenham - Petersham	27 Pittwater	43 Bringelly - Green Valley
12 Sydney Inner City	28 Warringah	44 Fairfield
13 Eastern Suburbs - North	29 Camden	45 Liverpool
14 Eastern Suburbs - South	30 Campbelltown (NSW)	46 Cronulla - Miranda - Caringbah
15 Bankstown	31 Wollondilly	47 Sutherland - Menai - Heathcote
16 Canterbury	32 Blue Mountains	

(b)



SA4 Name

1 Central Coast	6 Sydney - Inner South West	11 Sydney - Outer West and Blue Mountains
2 Sydney - Baulkham Hills and Hawkesbury	7 Sydney - Inner West	12 Sydney - Parramatta
3 Sydney - Blacktown	8 Sydney - North Sydney and Hornsby	13 Sydney - Ryde
4 Sydney - City and Inner South	9 Sydney - Northern Beaches	14 Sydney - South West
5 Sydney - Eastern Suburbs	10 Sydney - Outer South West	15 Sydney - Sutherland

Fig. 5. Digital boundaries: (a) SA3 level and (b) SA4 level.

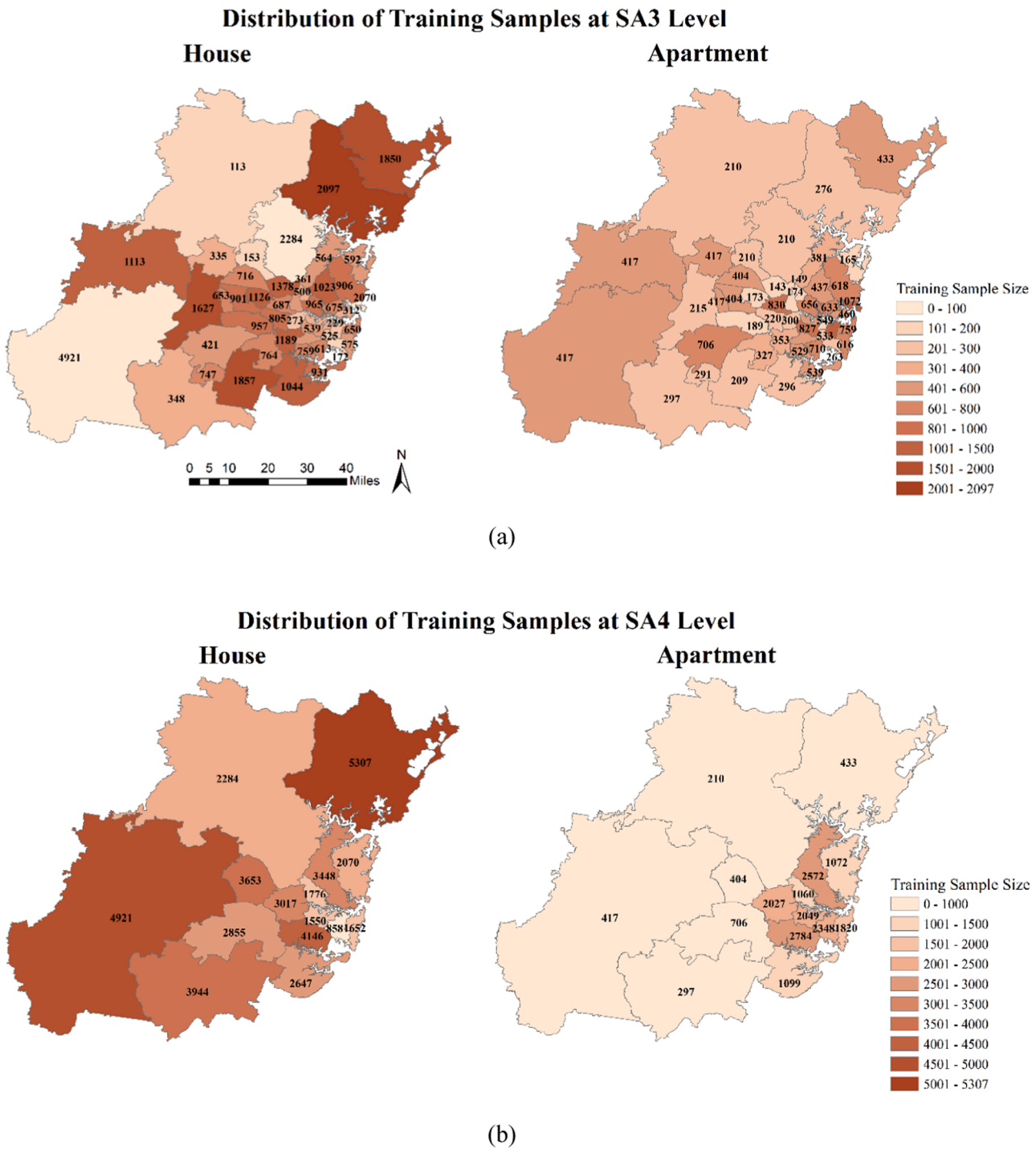


Fig. 6. The distribution of training samples for House and Apartment at: (a) SA3 level and (b) SA4 level.

4.2. Sampling strategy and parameter setting

The data sets are split into training and test sets prior to the experiments. We randomly selected 80 % samples as training data and the remaining as a test data set. All the results reported in this paper were averaged over 10 Monte Carlo runs with randomly selected training sets. During the experiments, the range of parameters is empirically determined, and the optimal values are determined by 5-fold cross-validation. The details about cross-validation technique can be referred to (Anguita et al., 2012; Burman, 1989; Wong and Yeh, 2019).

4.3. Data preprocessing

Data preprocessing has been done in this study prior to the modelling, including data cleaning, encoding categorical data, exploring feature correlation, and feature selection.

4.3.1. Data cleaning

Handling missing values and outliers are important for data cleaning. Attributes with large portion of missing value (more than 30 % missing) from original data have been removed, such as a set of property attributes such as swimming pool, solar panel, air-conditioning, heating, views and so on. In the experiments, the outliers were mainly removed

based on the extreme value of price and some key variables on the SA3 level. For each SA3 area, the sales record with extreme sales price and a set of other property features (e.g. number of bathrooms, number of bedrooms, number of parking, land size) were removed before training. The data is cleaned at 2.5 % and 97.5 % percentiles of some key variables to reduce the effects of extreme outliers. Any records that have a price less than 80 % of the 2.5th percentile or more than 120 % the 97.5th percentile are removed.

After data cleaning for missing value and outliers, 44,128 (out of 54084) sales of houses and 19,298 (out of 22979) sales of units were remained for further analysis.

4.3.2. Encoding categorical data

The categorical variable, which is property type in this study has been encoding to numerical variable before modelling.

4.3.3. Feature selection

With details of all input data outlined in previous sections, an initial list of variables were measured. The final list of variables was chosen following a number of reasons including: (i) they have an anticipated contribution to property price variance, having been found to be statistically significant in other Sydney land valuation studies (Mulley et al., 2016) and (ii) they are not associated with serious multicollinearity issues. Pearson's pairwise correlation coefficients and variance inflation factors (VIF) were used to detect any multicollinearity between variables. Variable pairs with a Pearson's correlation coefficient above 0.7 and a VIF above 4 were investigated, and some variables were omitted to ensure that the model outputs are not impaired by serious multicollinearity issues. For example, the percentage of professionals is highly correlated with both the family weekly income and percentage of people unemployed, and distance to CBD which is highly correlated with distance to major city centers, were removed from the modelling. The descriptive statistics for variables included in the property valuation models are provided in Table 3.

After the preprocessing of the dataset, the average price of a House and Apartment (Unit) is $1190,538 and $806,312, and the standard deviation of the House and Apartment (Unit) is 866,368 and 420,647, respectively. In total there are 23 features are used in the modelling process. The features are normalized to [0,1] before inputting into the models.

4.4. Validation approach

With respect to model validation, several accuracy metrics have been used. These are namely the R^2 score, Percentage Predicted Error within 5 %, 10 % and 20 %, respectively (PPE5, PPE10 and PPE20), Mean Absolute Percentage Error (MAPE), and Median Absolute Percentage Error (MdAPE).

R^2 is known as the coefficient of determination and denotes the proportion of the variance in the dependent variables (e.g. housing prices) that is predictable from the independent variable (e.g. housing features). The higher R^2 indicates a better predictive ability of the models.

PPEs validate the models in terms of a combination of accuracy and precision. The PPE is the percentage of properties of which the prices have been predicted by a model within a + /- percentage (e.g. 5 %, 10 % and 20 %). The higher PPEs, the more accurate and precise the model.

MAPE measures the accuracy of valuation models, and it is expressed as a percentage. For example, a MAPE value of 15 % means that the average difference between the predictions and the expected value is 15 %.

MdAPE calculates the median of the absolute percentage error. A 10 % MdAPE means that half of the absolute percentage errors are less than 10 %, and the others are larger than 10 %.

The calculation of each score is shown in Table 4.

Table 3
The descriptive statistics for variables.

Variable Name	mean	min	max	std
Sold Price (AUD)	1171,391	304,000	17,000,000	856,618.97
Number of Bedrooms (numerical variable)	3.588258	1	7	0.84
Number of Bathrooms (numerical variable)	1.845765	1	6	0.74
Number of Garages (numerical variable)	1.946238	0	15	0.95
Property Type (catergorial variable)	1.329299	0	5	0.94
Area Size (m^2)	866.2434	20	105400	1551.17
Longitude (decimal degree)	151.0359	150.25	151.5903	0.24
Latitude (decimal degree)	-33.7836	-34.32	-33.1319	0.21
Distance to City Center (m)	15191.11	144.11	58787.25	14146.85
Median Weekly Family Income (AUD)	2102.325	0	5045	677.72
Distance to Railway Station (m)	2972.679	39.68	27688.94	2819.23
Distance toTransmission Line (m)	1976.94	0.14	14886.87	2591.15
Percentage of Population Aged 65 above (numerical variable)	13.02855	0	93.53448	7.87
Distance to Coast (m)	22134.44	17.77	98914.35	17291.55
Distance to Swimming Pool (m)	2567.635	56.90	22065.81	1788.93
Distance to High School (m)	1451.475	36.51	25830.5	1483.29
Crime Rate (numerical variable)	0.0734	0.0079	1.483853	0.10
Distance to University (m)	7202.936	94.97	46993.44	6235.23
Percentage of Populatipon Born Overseas (numerical variable)	31.975	0	89.46237	13.93
Distance to Shopping Center (m)	2243.508	39.7	35067.61	3211.46
Distance to Hospital (m)	3611.332	57.20	31139.32	2961.46
Distance to Primary School (m)	693.4038	27.87	10698.63	455.97
Distance to Main Road (m)	547.6633	0	13581.45	495.94

* For property type: Cottage = 0, House = 1, Semi = 2, Terrace = 3, Townhouse = 4, Villa = 5.

Table 4
Equations for the calculation of validation metrics.

Name	Equation
PE	$PE = \left\|\frac{\widehat{y}_i - y_i}{y_i}\right\| * 100\%$
MAPE	$MAPE = mean(\|PE_i\|)$
MdAPE	$MdAPE = median(\|PE_i\|)$
PPE5/PPE10/PPE20	$PPE5 = ((PE <= 0.05).sum())/len(PE) * 100$ $PPE10 = ((PE <= 0.1).sum())/len(PE) * 100$ $PPE20 = ((PE <= 0.2).sum())/len(PE) * 100$

5. Experimental results and analysis

5.1. Model accuracy

Table 5 gives a summary of in-time validation results of multiple models at the three different geographical scales Overall, ML algorithms (except for SVR) and and GWR outperformed the linear regression approaches. The highest R^2, PPE5, PPE10, PPE20 and lowest MAPE and MdAPE were obtained by GBM, which are 90.99, 57.11, 75.63, 91.25, 7.38 and 3.96, respectively. For linear regression techniques, OLS and Ridge resulted in better performance and robustness than Lasso and Elastic Net in the experiments.while Random Forest, GBM, XGBoost, and MLP yielded similar results which have higher accuracy with a relative R^2 of 10 %. It should be noted that the performance of algorithms also relies on parameter selection, variation of the parameters may lead to different results. However, the results are likely to have a small

Table 5
Accuracies (%) of different property value models for houses.

Method	Sub-area Level	R^2	PPE5	PPE10	PPE20	MAPE	MdAPE
GWR	GSYD	**84.72**	**32.09**	**55.80**	**83.78**	**11.32**	**8.56**
OLS	SA3	**81.79**	**27.10**	**50.23**	**78.25**	**13.29**	**9.93**
	SA4	81.16	25.07	46.56	75.03	14.25	10.91
	GSYD	71.49	18.85	37.49	65.14	17.70	13.88
Lasso	SA3	**66.90**	**18.82**	**37.24**	**63.67**	**18.85**	**14.35**
	SA4	63.86	17.53	33.56	59.78	20.25	15.79
	GSYD	50.37	12.69	25.02	48.36	25.56	20.80
Ridge	SA3	**81.36**	**26.50**	**49.95**	**77.52**	**13.47**	**10.02**
	SA4	81.13	24.91	46.29	75.03	14.29	10.95
	GSYD	71.48	18.86	37.48	65.16	17.70	13.86
ElasticNet	SA3	**66.90**	**18.82**	**37.22**	**63.71**	**18.85**	**14.37**
	SA4	63.86	17.59	33.56	59.80	20.25	15.79
	GSYD	50.36	12.75	24.97	48.38	25.57	20.79
SVR	SA3	**68.63**	**19.55**	**37.93**	**64.90**	**18.26**	**13.78**
	SA4	64.56	17.91	35.63	62.18	19.32	15.08
	GSYD	55.37	17.00	32.78	57.67	21.77	16.53
Decision Tree	SA3	**85.14**	**36.16**	**59.39**	**83.99**	**11.34**	**7.81**
	SA4	84.13	32.17	55.88	81.79	12.13	8.49
	GSYD	76.02	23.93	44.77	72.58	15.32	11.47
Random Forest	SA3	**90.87**	**51.51**	73.98	91.16	**7.81**	4.80
	SA4	90.79	51.37	**74.24**	91.01	**7.81**	**4.78**
	GSYD	90.56	**51.51**	73.92	**91.19**	7.85	4.79
GBM	SA3	**90.99**	**57.11**	**75.63**	**91.25**	**7.38**	**3.96**
	SA4	90.17	51.05	73.65	91.05	7.96	4.84
	GSYD	87.98	38.20	64.87	88.35	9.67	6.84
XGBoost	SA3	**90.02**	**37.52**	**63.49**	**88.69**	**9.68**	7.17
	SA4	89.89	37.06	63.28	88.15	9.79	**7.14**
	GSYD	87.35	32.96	59.27	85.10	10.84	8.01
MLP	SA3	**89.02**	**37.26**	**63.41**	**88.40**	**9.85**	**7.12**
	SA4	87.41	33.87	60.13	86.39	10.66	7.84
	GSYD	83.40	29.31	53.40	80.33	12.49	9.16

difference as we selected the optimal parameters by using cross-validation during the experiments.

The prediction accuracies for Apartments (Unit) are shown in Table 6. It could be seen that results are consistent with models for Houses, where ML methods obtained better performance than the linear regression models. The best results were also generated by the GBM,

Table 6
Accuracies (%) of different property value models for apartments.

Method	Sub-area Level	R^2	PPE5	PPE10	PPE20	MAPE	MdAPE
GWR	GSYD	**79.55**	**30.10**	**55.18**	**82.95**	**11.64**	**8.81**
OLS	SA3	**72.95**	**31.50**	**56.73**	**82.72**	**11.66**	**8.51**
	SA4	69.09	27.14	52.36	80.39	12.62	9.42
	GSYD	62.84	23.37	43.48	73.28	14.97	11.86
Lasso	SA3	**41.92**	**19.48**	**37.49**	**64.32**	**19.02**	**14.01**
	SA4	36.00	16.20	32.11	59.41	20.83	16.06
	GSYD	33.68	14.47	30.03	55.27	22.17	17.67
Ridge	SA3	**69.86**	**29.86**	**52.71**	**80.52**	**12.45**	**9.33**
	SA4	65.52	26.26	48.74	77.98	13.44	10.32
	GSYD	62.83	23.29	43.42	73.30	14.97	11.83
ElasticNet	SA3	**41.92**	**19.42**	**37.41**	**64.38**	**19.02**	**14.02**
	SA4	36.01	16.26	32.04	59.51	20.84	16.07
	GSYD	34.34	14.41	30.28	55.87	22.06	17.68
SVR	SA3	**43.90**	**21.29**	**39.49**	**66.40**	**17.97**	**13.29**
	SA4	38.97	18.36	35.75	62.69	19.25	14.49
	GSYD	35.67	18.13	34.48	60.66	19.90	15.68
Decision Tree	SA3	**77.08**	**43.50**	**66.50**	**87.32**	**9.89**	**6.05**
	SA4	76.86	37.35	62.55	85.69	10.60	7.03
	GSYD	69.10	28.59	52.94	80.06	12.83	9.26
Random Forest	SA3	86.79	53.38	76.52	92.49	7.38	4.52
	SA4	**86.85**	53.21	76.48	92.56	7.37	4.53
	GSYD	86.81	**55.44**	**77.46**	**93.18**	**7.12**	**4.31**
GBM	SA3	**86.66**	**56.12**	**77.00**	**92.26**	**7.23**	**4.17**
	SA4	85.62	53.21	76.23	92.18	7.50	4.56
	GSYD	82.39	43.86	72.09	91.10	8.38	5.87
XGBoost	SA3	**85.94**	38.18	**66.29**	89.21	9.27	6.78
	SA4	85.73	**38.49**	66.19	**89.94**	**9.21**	**6.73**
	GSYD	83.60	36.91	64.71	89.25	9.56	7.05
MLP	SA3	**84.89**	**42.57**	**69.62**	**90.60**	**8.77**	**6.07**
	SA4	82.80	38.64	65.40	88.81	9.53	6.81
	GSYD	77.12	31.42	57.89	85.34	11.07	8.25

* The best results were highlighted in bold.

however, the differences between this approach and Random Forest, MLP, and XGBoost are not significant, which indicates that those methods have a competitive and robust capability in predicting the property values.

When comparing the performance of models across different sub-areas, it could be seen that models with the most granular geographic scale - SA3 reached the highest accuracy, the superiority of smaller geographic units is distinct in linear regression, where the PPE20 of house models running in SA3 areas are almost 15 % higher than those running across the whole city. And it should be noted that the local spatial regression GWR outperformed the linear regression run at sub areas, this could be associated to two reasons: 1. The capability of dealing the spatial dependence issue in GWR. 2. SA3 is not granular enough to capture the local markets. Almost all the Ml algorithms show better results for models running at sub areas than at city scale. However, there is no significant difference between the results obtained on SA3 and SA4 levels which implies that ML algorithms have better capability in capturing geographical discriminative information from the various input attributes. It should be noted that Random Forest model performed similar at SA3, SA4 and city scale which indicates that it is less prone to spatial dependency issue.

5.2. Impacts of hedonic attributes

In order to understand the importance of different hedonic attributes in property value prediction and also to examine impacts of attributes across areas, the feature importance scores were illustrated in this study. SHAP (Lundberg and Lee, 2017) has been employed in this study to help display the impact of each variable on housing price with strength and direction, and it is available on https://github.com/slundberg/shap.

It has been observed from the experiments, the feature importance of the applied ML models in this study are similar, and here Random Forest is taken as an example. Fig. 7 shows the feature importance of Random Forest model as city scale for House. The figure shows property structural features, such as number of bedrooms, number of bathrooms and land size, are more influential in the model, which is quite consistent with the results from linear regression models in the past studies (Mulley et al., 2016; Pettit et al., 2020). It could be seen that the geographic location (captured by longitude and latitude) also plays important roles in property price prediction, which indicates that prices vary across space and spatial dependence might exist in house prices. Socio-economic level is another important predictor in house prices, measured by the median family income in neighbourhood, which is under expectation and also consistent with previous studies (Goodman and Thibodeau, 2003; Pettit et al., 2020).

The feature impacts are different on different SA levels. Fig. 8 shows the feature importance for House in the suburb of Miranda at SA3 (Cronulla – Miranda - Caringbah), and SA4 (Sutherland) levels. It can be observed from Figs. 7–8 that various features show different importance at different SA levels, including GSYD. At the GSYD level, area size is the fourth most important factor contributing to the results, however, the property type and locational attributes contribute more than area size at a smaller regional scale. At the SA3 level, property type is more important than at SA4 and GSYD level. This may be due to locational attributes are similar at a small scale, therefore property type is considered more important. However, this is not consistent for all sub-areas, such as in Carlingford (SA3), Auburn (SA3) and their corresponding SA4 area - Paramatta. As shown in Fig. 9, area size has a similar importance for Houses locate in these three areas, it may because that Paramatta is relative far away from the city center and Houses in this region usually are situated on large blocks of land. Given a larger enough block, the design of bedrooms and bathrooms is more flexible, so the area size matters more in housing price. However, within the same SA4 area, the important influential factors are different for Carlingford and Auburn.

It has also been noted that the distance to city center and shopping center sometimes show positive impacts on the valuation in local areas, which is contrary to common sense. In such areas, accessibility may be not an issue. The negative effects of distance factors on property values are more likely to be driven by the spatial pattern that big houses usually locate far from the city center. It is similar for the impact of distance to railway station on property valuation in local areas, as the railway stations always locates in the city centers in Sydney. This also may due to the gentrification is expanding outwards from the city center in Sydney in recent years. The chosen examples in this study are the gentrification hotspots in the western and southern in Sydney, respectively. The rapid gentrification and suburbanization make the relationship between housing price and distance to city center is not simple anymore. Analysis has characterized the suburbs in Australia are becoming more fragmented and heterogenous (Baum et al., 2005).

Geographical location and neighbourhood features play significant roles in determining property values. For example, properties in North Sydney are more expensive than surrounding areas because they have access to high quality public schools, which also reflected in the feature importance as shown in Fig. 10.

The feature importance for Apartment (Unit) is similar to the ones for House. However, the analysis may be limited by the data availability, such as property age, year of refurbishment, wall material, energy

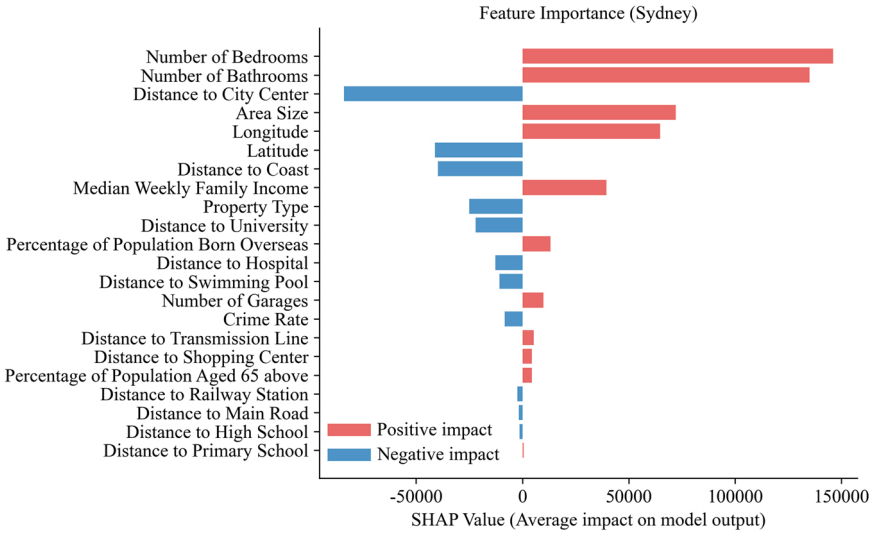


Fig. 7. The illustration of the feature importance of Random Forest model for House in Sydney.

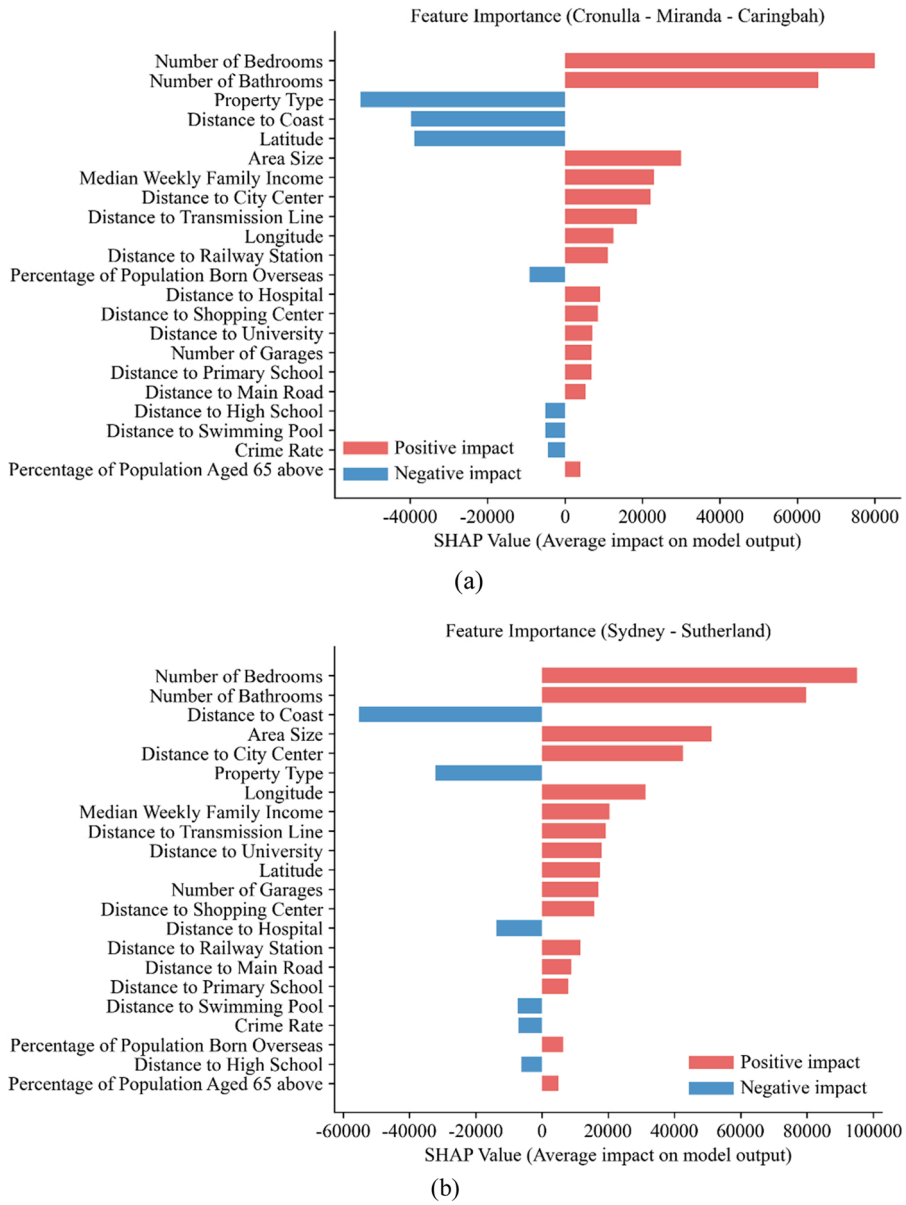


Fig. 8. Feature importance for House in Miranda at SA3 and SA4 levels: (a) Miranda and (b) Sutherland.

ratings which may also influence the property price.

5.3. Impact of the number of threshold samples

In this section, we test the impact of the number of threshold training samples on the accuracies for both data sets, and Random Forest is taken as an example. In the experiment, the threshold training samples as 20, 50, 100, 150 and 200, and the samples are selected randomly. The validation results with the increase of the threshold training samples are shown in Fig. 11. From the figure, it can be seen that the best results have been achieved with the number is set as 100 and 150 for House and Apartment, respectively. With a smaller number, the model is more likely developed at a smaller SA level, however, the prediction capability is comprised due to the limited training samples. On contrast, the model is more likely developed at the upper SA level, in which situation the local geographical attributes may not be characterized. Although the results are similar with the number is setting as 100 and 150, 100 has been chosen in this paper given that the number of sold properties in the SA3 level may not be sufficient in real world.

5.4. Out-of-time prediction accuracy

In real property market, another important indicator for evaluating model performance is the out-of-time forecast validation as people want to predict future housing prices. In this study, we also examine the capability of various algorithms in predicting future property price. The transaction records for Quarter 1 of 2019 was used as the validation samples. As observed from Section 5.2, most of models achieved best results at SA3 level, therefore the parameters were chosen for the models with the SA3 level in this section. Tables 7–8 show the out-of-time forecast evaluation results for each model for House and Apartment (Unit), respectively. It can be seen that models for houses are less performed than the out-of-sample validation, where the PPE 5 from the best model decreased from around 52–33 %. While the performance of out-of-time prediction for apartments are similar to the out-of-sample validation. The result indicates that some price adjustments need to be considered when applying the AVM in real time, particulary for houses, where the market is much more volatile than apartments.

Figs. 12–13 illustrates the model performance in terms of R^2 and PPE20, as well as the frequently applied indicator in industry - hit rate

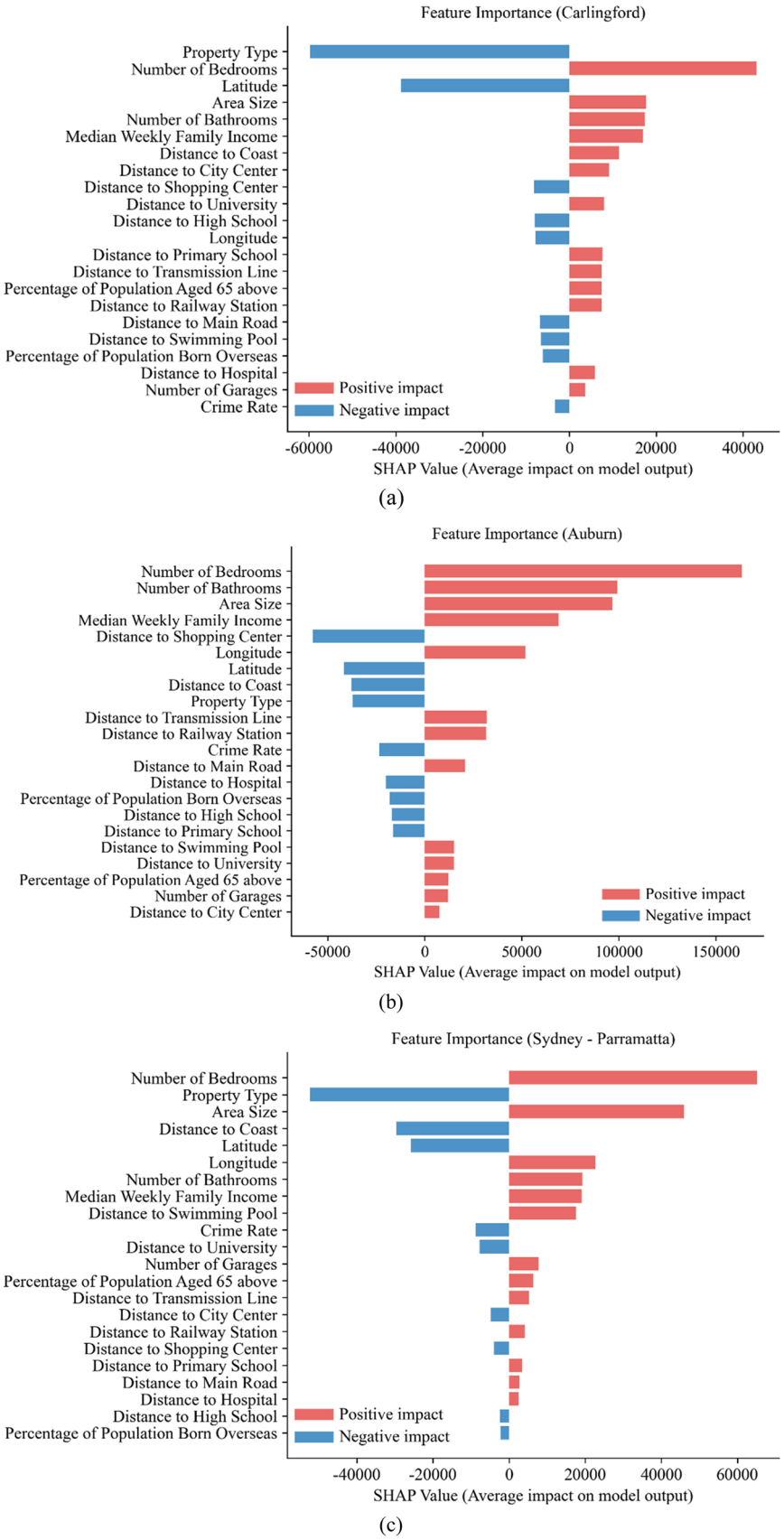


Fig. 9. Feature importance for House in Parramatta at SA3 and SA4 levels: (a) Carlingford, (b) Auburn, (c) Parramatta.

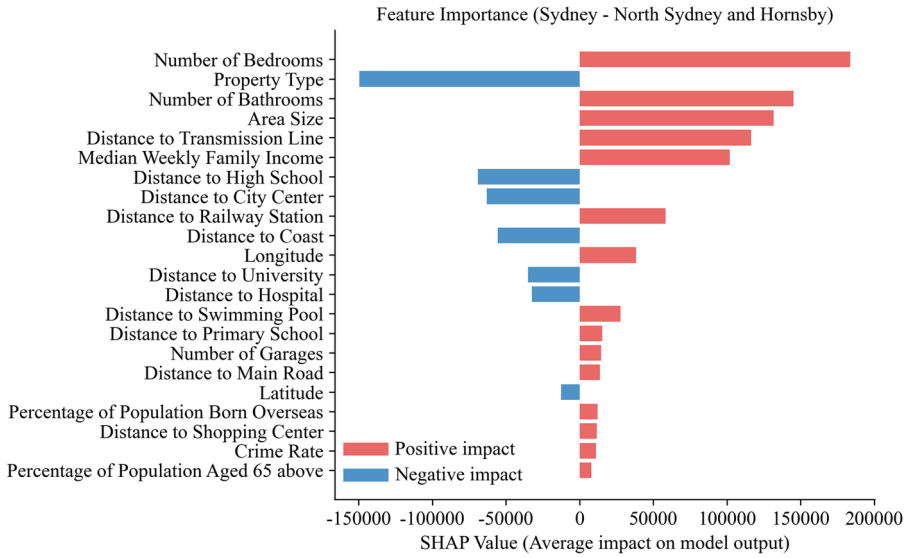


Fig. 10. Feature importance for House in North Sydney and Hornsby.

(PPE10) with MAPE (CoreLogic, 2018).

The results from Tables 7–8 and Figs. 12–13 also imply that the Random Forest and Gradient Boosting-based models perform better than other ML models, as they have higher R^2, PPEs and lower MAPE and MdAPE. Even though GBM and Random Forest have the best R^2 for House and Apartments in terms of out-of-time validation (87.90 and 82.34, respectively), their performance are not comparable with the in-time evaluation. The limited number of transaction records may be one of the possible reasons for the lower performance as there are more observations for in-sample validation. Another reason is that the models are selected based on the performance of the in-time validation data. This finding suggests that using the most recent training data may improve the prediction results if given sufficient training samples.

6. Conclusion

In this paper, several branches of ML methods were applied for predicting property valuation. In support of the modelling, we collected a rich set of data from a diverse range of sources, which is necessary in order to formulate accurate predictions. In terms of property type, both House and Apartment are valued. In addition, the application was carried out on three different geographical areas - statistical area levels, i.e. SA3, SA4, and GSYD to understand different housing markets across the metropolitan area. Considering the practical problem of limited training samples in small areas, a multi-level strategy was developed. Dwelling attributes, locational, and statistical factors are included as explanatory variables in modelling with the sold price was selected as the dependent variable.

Experiments were conducted on the housing market of Greater Sydney, and various accuracy metrics were applied in this paper to validate the performance. Not only the frequently used indicators as reported in the literature, such as R^2, MAPE, and MdAPE, but also the reference scores in the real-world property industry, i.e. PPEs, were used to evaluate the modelling performance. The results show that ML algorithms and GWR outperform some of the traditional HPMs (e.g. OLS, Lasso, and Ridge). Random Forest, XGBoost, Gradient Boosting, and MLP provide the best performance in terms of the majority of the validation approaches. Moreover, the dwelling attributes for House were shown to contribute to most substantially improving the performance of the property prediction models. Overall, among all the considered factors which contributed most to the models including the number of bedrooms, area size, and distance to city center. In contrast, the number of garages, distance to the main road, and distance to schools contribute least to determining the predicted property value. This study also shows the contribution of each factor in different local makerts, which is insightful in providing suggestions for policy-making.

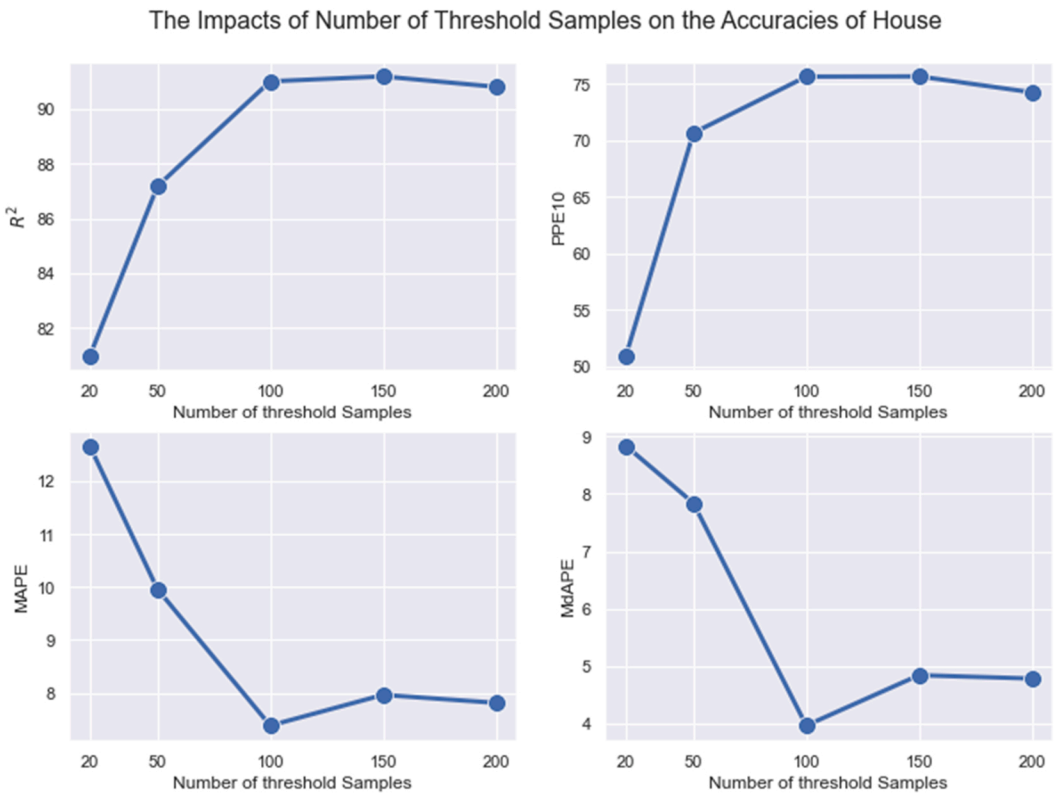
Both in-time and out-of-time validation were implemented in this study. As might be expected, the findings suggest that the most recent training data can maximize the predictive capability of the models, and the number of training samples would influence the modelling performance. In terms of the application of statistical area levels, results indicate that valuation carried on smaller regions may improve the predictive performance of most of the modelling techniques. In this research, models perform best on SA3 level, and perform better on SA4 level than City level. Interestingly the traditional regression methods have gained more improvements of performance than machine learning methods when carried on smaller housing markets.

In this research, various performance validation metrics are applied, including R^2, PPEs, MAPE and MdAPE, which have been frequently used in literature. Especially, for out-of-time validation, Hit Rate which is widely applied in Industry are also included in this paper.

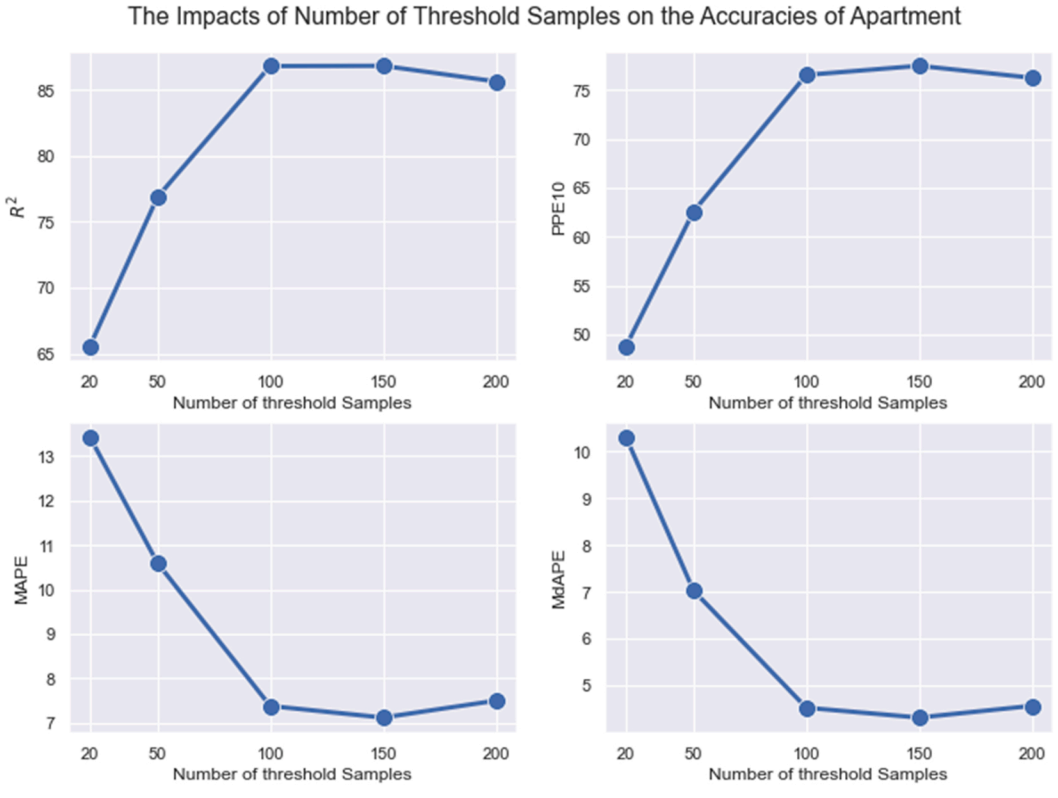
In addition to the methodological contributions, our findings by property valuation provide evidence informing policy makings for equable and sustainable urban development. First, locational factors (captured by longitude and latitude) play a vital role in determining property values, implying an uneven distribution across the Greater Sydney region. The wealthy people prefer to reside in the northern and eastern parts of the region, while low-cost housing mainly concentrates in the western and northern areas. The spatial distribution pattern is especially significant for the Sydney city given the greatest importance of locations. This finding echos the recent trend of suburb poverty in the Western Sydney (Bangura and Lee, 2019), against the development of diverse and inclusive city.

Moreover, accessibility factors (i.e., distance to city center, distance to coast, distance to university and distance to hospital) are identified to significantly influence property values, particularly in Sydney city. Sorting by housing, the advantaged groups have higher accessibility to employment opportunities and leisure, education, medical resources, suggesting the socio-spatial inequalities. The reinforcing effect of accessibility on housing differentiation needs to be paid more attention when implementing welfare regimes related to housing policies.

In contrast to the Sydney city, in some small suburb areas like Miranda, property value is primarily determined by the housing characteristics, such as the number of bedrooms, the number of bathrooms and housing size. In such areas, accessibility may not be an issue. The negative effects of distance factors on property values are more likely to



(a)



(b)

Fig. 11. The impact of the number of threshold samples on the prediction accuracies: (a) House, (b) Apartment.

Table 7
Out-of-time validation results of various models for House.

Method	R^2	PPE5	PPE10	PPE20	MAPE	MdAPE
GWR	84.99	23.53	43.98	74.71	14.73	11.69
Lasso	60.93	17.82	35.1	60.67	20.15	16.37
Ridge	76.61	25.5	41.34	70.3	15.65	12.51
ElasticNet	61.23	16.82	32.37	58.71	20.13	16.3
OLS	74.54	25.1	40.23	70.68	14.26	11.25
SVR	58.95	16.95	32.93	59.89	20.37	14.35
Decision Tree	80.14	31.21	51.68	76.93	14.46	**8.98**
Random Forest	86.21	31.59	56.72	83.24	12.78	9.63
GBM	**87.90**	32.78	**57.7**	84.69	13.41	9.43
XGBoost	86.31	**32.88**	56.21	**85.1**	12.7	9.04
MLP	84.23	31.26	53.41	84.47	**11.86**	8.99

Table 8
Out-of-time validation results of various models for Apartment (Unit).

Method	R^2	PPE5	PPE10	PPE20	MAPE	MdAPE
GWR	72.94	24.75	47.25	77.50	13.85	10.75
Lasso	31.64	13.15	30.14	54.25	22.78	18.22
Ridge	61.34	22.13	43.21	74.32	15.55	12.22
ElasticNet	30.56	14.01	29.68	54.27	23.12	18.36
OLS	60.24	23.24	40.56	72.11	15.2	12.45
SVR	34.12	18.01	32.64	61.2	19.44	15.68
Decision Tree	66.32	28.63	51.67	79.22	13.52	9.97
Random Forest	**82.34**	**54.99**	**76.54**	**90.12**	**8.66**	**4.67**
GBM	80.17	43.56	72.15	88.98	9.13	5.46
XGBoost	81.67	35.67	64.25	88.12	9.23	7.06
MLP	78.12	31.67	58.98	84.22	11.1	8.35

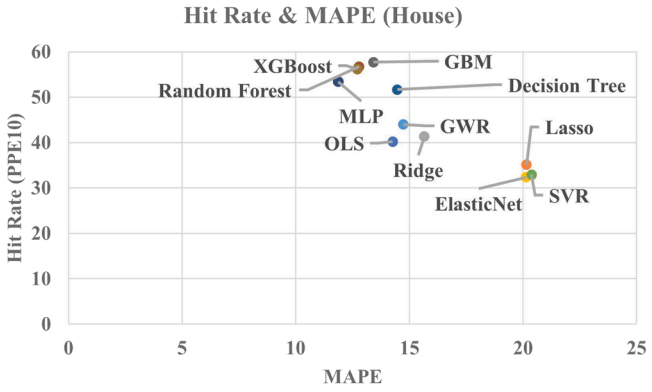


Fig. 12. Model performance for out-of-time validation of House.

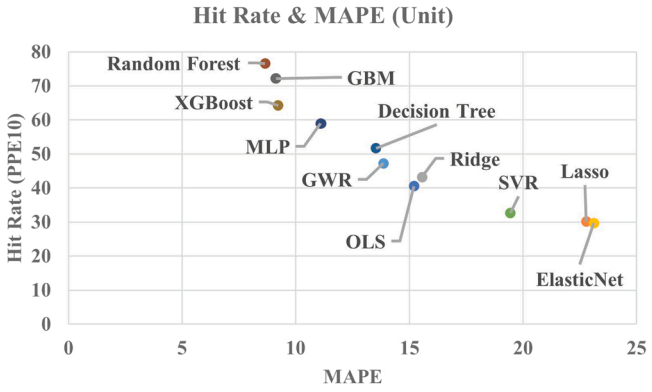


Fig. 13. Model performance for out-of-time validation of Apartment (Unit).

be driven by the spatial pattern that big houses usually locate far from the city center. Unlike the Sydney city, where lots of people still rely on public transit, people in suburbs usually travel by private car for long-distance trips and walking for nearby opportunities and activities, thus are less sensitive to distance. In this regard, influential factors of property value in different local areas depend on geographical contexts, such as city size and travel mode.

In addition, property valuations are helpful to understand the risks of gentrification by housing affordability. The ongoing increased housing price will deteriorate housing affordability of residents from low-income regions such as Western Sydney (Bangura and Lee, 2019). Under the housing crisis, people tend to move to affordable areas, and the disadvantaged groups will be most likely to be displaced by middle and high-income residents, exacerbating socio-spatial polarisation. A more accurate property evaluation can assist the government in making timely adjustments to decrease the risk of gentrification of low-income groups.

There are some data-driven limitations in the research which are important to discuss. For example, housing age is well-known as a significant factor impacting house price (Coulson and McMillen, 2008; Palm et al., 2020) as well as building internal area and floor level (for units). However, due to the data not being comprehensibly available for the study area, those attributes could not be incorporated in the current models. Another data limitation is that all the accessibility variables in this paper are measured by point-to-point distance, while the size and functional scales of facilities (e.g. stations, schools, and hospitals) which might vary across the city were not considered in this study due to data gaps. This could be future efforts to improve the predictive accuracy.

Another limitation is that the models were implemented on each SAs individually, which ignores the relation among each sub-area (Randolph and Tice, 2013). Some research has been undertaken from the perspective of using submarkets to characterize the neighbourhood effects on the house price. The importance of the submarket in the public housing sector has been long investigated since the 1970 s (Grigsby, 1986, 1963). However, the definition of a housing submarket has never been straightforward (Bourassa et al., 1999; Grigsby, 1986). In the research (Bangura and Lee, 2020; Costello et al., 2019; Randolph and Tice, 2014), the authors identified that social, spatial factors and socio-economic factors would influence in defining housing submarket differently. In future, we will conduct analysis based on clustering techniques which are helpful in grouping similar submarkets by considering both statistical and socio-economic factors (Bangura and Lee, 2021), so that submarkets can be defined more reasonable and the determinants of housing price can be studied more comprehensively.

Code availability

The codes are available by contacting the authors.

Conflict of interest

No competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv Prepr. arXiv 1603, *04467*.
Alpaydin, E., 2020. Introduction to Machine Learning. MIT press.
Anderson, J.A., 1995. An Introduction to Neural Networks. MIT press.
Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S., 2012. The'K'in K-fold Cross Validation. ESANN.

Anselin, L., 2013. Spatial econometrics: methods and models, vol. 4. Springer Science & Business Media.
Antipov, E.A., Pokryshevskaya, E.B., 2012. Mass appraisal of residential apartments: an application of Random forest for valuation and a CART-based approach for model diagnostics. Expert Syst. Appl. 39 (2), 1772–1778.
Armstrong, R.J., Rodriguez, D.A., 2006. An evaluation of the accessibility benefits of commuter rail in eastern Massachusetts using spatial hedonic price functions. Transportation 33 (1), 21–43.
Awad, M., Khanna, R., 2015. Support vector regression. Efficient learning machines. Springer, pp. 67–80.
Azmoodeh, M., Haghighi, F., Motieyan, H., Tilaki, M.J.M., 2020. Investigating the relationship between housing policy and accessibility, based on developing a multi-perspectives accessibility index: a case study in Tehran, Iran. J. Hous. Built Environ. 1–23.
Bangura, M., Lee, C.L., 2019. The differential geography of housing affordability in Sydney: a disaggregated approach. Aust. Geogr. 50 (3), 295–313.
Bangura, M., Lee, C.L., 2020. House price diffusion of housing submarkets in Greater Sydney. Hous. Stud. 35 (6), 1110–1141.
Bangura, M., Lee, C.L., 2021. The determinants of homeownership affordability in Greater Sydney: evidence from a submarket analysis. Hous. Stud. 1–27.
Baum, S., O'Connor, K., Stimson, R., 2005. Fault lines exposed: advantage and disadvantage across Australia's settlement system. Monash Univ. ePress.
Bento, A., Lowe, S., Knaap, G.-J., Chakraborty, A., 2009. Housing market effects of inclusionary zoning. Cityscape 7–26.
Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25 (2), 197–227.
Bin, O., Landry, C.E., 2013. Changes in implicit flood risk premiums: empirical evidence from the housing market. J. Environ. Econ. Manag. 65 (3), 361–376.
Bourassa, S., Cantoni, E., Hoesli, M., 2010. Predicting house prices with spatial dependence: a comparison of alternative methods. J. Real. Estate Res. 32 (2), 139–160.
Bourassa, S.C., Hamelink, F., Hoesli, M., MacGregor, B.D., 1999. Defining housing submarkets. J. Hous. Econ. 8 (2), 160–183.
Bourassa, S.C., Hoesli, M., Merlin, L., Renne, J., 2020. Big data, accessibility, and urban house prices. Urban Stud.
Buonanno, P., Montolio, D., Raya-Vílchez, J.M., 2013. Housing prices and crime perception. Empir. Econ. 45 (1), 305–321.
Burman, P., 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika 76 (3), 503–514.
Case, B., Clapp, J., Dubin, R., Rodriguez, M., 2004. Modeling spatial and temporal house price patterns: a comparison of four models. J. Real. Estate Financ. Econ. 29 (2), 167–191.
Case, K.E., Mayer, C.J., 1996. Housing price dynamics within a metropolitan area. Reg. Sci. Urban Econ. 26 (3–4), 387–407.
Čeh, M., Kilibarda, M., Lisec, A., Bajat, B., 2018. Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS Int. J. Geo-Inf. 7 (5), 168.
Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2 (3), 1–27.
Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,
Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. R package version 0.4–2, 1(4).
Chen, X., Wei, L., Xu, J., 2017. House price prediction using lstm. arXiv Prepr. arXiv 1709, 08432.
Chica-Olmo, J., Cano-Guervos, R., Chica-Rivas, M., 2019. Estimation of housing price variations using spatio-temporal data. Sustainability 11 (6), 1551.
Chung-Ang, 2019. Comparing the housing implicit prices of restricted and unrestricted hedonic price models. J. Korea Plan. *J. Korea Plan. Assoc.* 54 (6), 80–88.
CoreLogic. (2018). Residential Property Index Series.⟨https://www.corelogic.com.au/sites/default/files/2018–01/Residential-Property-Index-Series.pdf⟩.
Costello, G., Leishman, C., Rowley, S., Watkins, C., 2019. Drivers of spatial change in urban housing submarkets. Geogr. J. 185 (4), 432–446.
Coulson, N.E., McMillen, D.P., 2008. Estimating time, age and vintage effects in housing prices. J. Hous. Econ. 17 (2), 138–151.
Dai, X., Bai, X., Xu, M., 2016. The influence of Beijing rail transfer stations on surrounding housing prices. Habitat Int. 55, 79–88.
Diaz, R.B., Mclean, V., 1999. Impacts of rail transit on property values. Am. Public Transit Assoc. Rapid Transit Conf. Proc.
Du, H., Mulley, C., 2006. Relationship between transport accessibility and land value: local model approach with geographically weighted regression. Transp. Res. Rec. 1977 (1), 197–205.
Du, H., Mulley, C., 2012. Understanding spatial variations in the impact of accessibility on land value using geographically weighted regression. J. Transp. Land Use 5 (2), 46–59.
Dziauddin, M.F., Powe, N., Alvanides, S., 2015. Estimating the effects of light rail transit (LRT) system on residential property values using geographically weighted regression (GWR). Appl. Spat. Anal. Policy 8 (1), 1–25.
Englund, P., Ioannides, Y.M., 1997. House price dynamics: an international empirical perspective. J. Hous. Econ. 6 (2), 119–136.
Evangelio, R., Hone, S., Lee, M., Prentice, D., 2019. What makes a locality attractive? Estimates of the amenity value of parks for Victoria. Econ. Pap.: A J. Appl. Econ. Policy 38 (3), 182–192.
Fan, C., Cui, Z., Zhong, X., 2018. House prices prediction with machine learning algorithms. Proc. 2018 10th Int. Conf. Mach. Learn. Comput.
Fan, G.-Z., Ong, S.E., Koh, H.C., 2006. Determinants of house price: a decision tree approach. Urban Stud. 43 (12), 2301–2315.
Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008. LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874.
Farlow, A., 2005. UK house prices, consumption and GDP in a global context. Memo, Department of Economics and Oriel College. University of Oxford.
Feuillet, T., Commenges, H., Menai, M., Salze, P., Perchoux, C., Reuillon, R., Kesse-Guyot, E., Enaux, C., Nazare, J.-A., Hercberg, S., 2018. A massive geographically weighted regression model of walking-environment relationships. J. Transp. Geogr. 68, 118–129.
Fik, T.J., Ling, D.C., Mulligan, G.F., 2003. Modeling spatial variation in housing prices: a variable interaction approach. Real. Estate Econ. 31 (4), 623–646.
Filippova, O., Sheng, M., 2020. Impact of bus rapid transit on residential property prices in Auckland, New Zealand. J. Transp. Geogr. 86, 102780.
Fotheringham, A.S., Brunsdon, C., Charlton, M., 2000. Quantitative Geography: Perspectives on Spatial Data Analysis. Sage.
Fotheringham, A.S., Brunsdon, C., Charlton, M., 2003. Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons.
Fotheringham, A.S., Crespo, R., Yao, J., 2015. Geographical and temporal weighted regression (GTWR). Geogr. Anal. 47 (4), 431–452.
Gao, G., Bao, Z., Cao, J., Qin, A.K., Sellis, T., Wu, Z., 2019. Location-centered house price prediction: a multi-task learning approach. arXiv Prepr. arXiv 1901, 01774.
Ge, C. (2019). A LSTM and Graph CNN Combined Network for Community House Price Forecasting. 2019 20th IEEE International Conference on Mobile Data Management (MDM),
Giglio, S., Maggiori, M., Stroebel, J., & Weber, A. (2015). Climate change and long-run discount rates: Evidence from real estate (0898–2937).
Glaeser, E.L., & Gyourko, J. (2002). The impact of zoning on housing affordability (0898–2937).
Goodman, A.C., Thibodeau, T.G., 2003. Housing market segmentation and hedonic prediction accuracy. J. Hous. Econ. 12 (3), 181–201.
Grigsby, W., 1986. The Dynamics of Neighborhood Change and Decline. Pergamon, London.
Grigsby, W.G., 1963. Housing Markets and Public Policy. University of Pennsylvania Press.
Haider, M., Miller, E.J., 2000. Effects of transportation infrastructure and location on residential real estate values: application of spatial autoregressive techniques. Transp. Res. Rec. 1722 (1), 1–8.
Hong, J., Choi, H., Kim, W.-S., 2020. A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. Int. J. Strateg. Prop. Manag. 24 (3), 140–152.
Huston, S., Lahbash, E., 2018. Land value capture and tax increment financing: overview and considerations for sustainable urban investment. Eur. J. Sustain. Dev. 2 (3).
Ihlanfeldt, K., Mayock, T., 2010. Panel data estimates of the effects of different types of crime on housing prices. Reg. Sci. Urban Econ. 40 (2–3), 161–172.
Jain, N., Goel, P., Sharma, P., & Deep, V. (2019). Prediction of House Pricing Using Machine Learning with Python. International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India,
Kendall, R., Tulip, P., 2018. The effect of zoning on housing prices. Reserve Bank Aust. Res. Discuss. Pap. (2018-03).
Kim, E.J., Kim, H., 2020. Neighborhood walkability and housing prices: a correlation study. Sustainability 12 (2), 593.
Lancaster, K.J., 1966. A new approach to consumer theory. J. Political Econ. 74 (2), 132–157.
Law, S., Paige, B., Russell, C., 2019. Take a look around: using street view and satellite images to estimate house prices. ACM Trans. Intell. Syst. Technol. (TIST) 10 (5), 1–19.
Lee, C.L., Locke, M., 2020. The effectiveness of passive land value capture mechanisms in funding infrastructure. J. Prop. Invest. Financ.
Li, H., Wei, Y.D., Wu, Y., Tian, G., 2019. Analyzing housing prices in Shanghai with open data: amenity, accessibility and urban structure. Cities 91, 165–179.
Liebelt, V., Bartke, S., Schwarz, N., 2019. Urban green spaces and housing prices: an alternative perspective. Sustainability 11 (13), 3707.
Limsombunchai, V., 2004. House price prediction: hedonic price model vs. artificial neural network. N. Z. Agric. Resour. Econ. Soc. Conf.
Löchl, M., Axhausen, K.W., 2010. Modeling hedonic residential rents for land use and transport simulation while considering spatial effects. J. Transp. Land Use 3 (2), 39–63.
Lu, S., Li, Z., Qin, Z., Yang, X., & Goh, R.S.M. (2017). A hybrid regression technique for house prices prediction. 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM),
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. *30*.
Mansfield, C., Pattanayak, S.K., McDow, W., McDonald, R., Halpin, P., 2005. Shades of green: measuring the value of urban forests in the housing market. J. For. Econ. 11 (3), 177–199.
Maser, S.M., Riker, W.H., Rosett, R.N., 1977. The effects of zoning and externalities on the price of land: an empirical analysis of Monroe County. N. Y. J. Law Econ. 20 (1), 111–132.
McDonald, G.C., 2009. Ridge regression. Wiley Interdiscip. Rev.: Comput. Stat. 1 (1), 93–100.
Mohd, T., Jamil, N.S., Johari, N., Abdullah, L., Masrom, S., 2020. An overview of real estate modelling techniques for house price prediction. Charting a Sustain. Future ASEAN Bus. Soc. Sci. 321–338.

Mohri, M., Rostamizadeh, A., Talwalkar, A., 2018. Foundations of Machine Learning. MIT press.
Mu, J., Wu, F., Zhang, A., 2014. Housing value forecasting based on machine learning methods. Abstr. Appl. Anal.
Mullainathan, S., Spiess, J., 2017. Machine learning: an applied econometric approach. J. Econ. Perspect. 31 (2), 87–106.
Mulley, C., Ma, L., Clifton, G., Yen, B., Burke, M., 2016. Residential property value impacts of proximity to transport infrastructure: an investigation of bus rapid transit and heavy rail networks in Brisbane, Australia. J. Transp. Geogr. 54, 41–52.
Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., 2004. An introduction to decision tree modeling. J. Chemom. Soc. 18 (6), 275–285.
Neill, H.R., Hassenzahl, D.M., Assane, D.D., 2007. Estimating the effect of air quality: spatial versus traditional hedonic price models. South. Econ. J. 1088–1111.
Noor, N.M., Asmawi, M.Z., Abdullah, A., 2015. Sustainable urban regeneration: GIS and hedonic pricing method in determining the value of green space in housing area. Procedia-Soc. Behav. Sci. 170, 669–679.
Pagliara, F., Papa, E., 2011. Urban rail systems investments: an analysis of the impacts on property values and residents' location. J. Transp. Geogr. 19 (2), 200–211.
Palm, M., Raynor, K.E., Warren-Myers, G., 2020. Examining building age, rental housing and price filtering for affordability in Melbourne, Australia. Urban Stud. 0042098020927839.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: machine learning in Python. J. Mach. Learning Res. 12, 2825–2830.
Peng, Z., Huang, Q., & Han, Y. (2019). Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm. 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT),
Pettit, C., Shi, Y., Han, H., Rittenbruch, M., Foth, M., Lieske, S., van den Nouwelant, R., Mitchell, P., Leao, S., Christensen, B., 2020. A new toolkit for land value analysis and scenario planning. Environ. Plan. B: Urban Anal. City Sci. 47 (8), 1490–1507.
Piao, Y., Chen, A., & Shang, Z. (2019). Housing Price Prediction Based on CNN. 2019 9th International Conference on Information Science and Technology (ICIST),
Rahman, S., Masih, M., 2014. Increasing household debts and its relation to GDP, interest rate and house price: Malaysia's perspective. MPRA. University Library of Munich, Germany.
Randolph, B., Tice, A., 2013. Who lives in higher density housing? A study of spatially discontinuous housing sub-markets in Sydney and Melbourne. Urban Stud. 50 (13), 2661–2681.
Randolph, B., Tice, A., 2014. Suburbanizing disadvantage in Australian cities: sociospatial change in an era of neoliberalism. J. Urban Aff., 36(sup1) 384–399.
Rojas, R., Feyen, L., Watkiss, P., 2013. Climate change and river floods in the European Union: socio-economic consequences and the costs and benefits of adaptation. Glob. Environ. Change 23 (6), 1737–1751.
Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. J. Political Econ. 82 (1), 34–55.
Rosewall, T., & Shoory, M. (2017). Houses and Apartments in Australia| Bulletin–June Quarter 2017. Bulletin(June).
Schulz, R., Werwatz, A., 2004. A state space model for Berlin house prices: estimation and economic interpretation. J. Real. Estate Financ. Econ. 28 (1), 37–57.
Se Can, A., Megbolugbe, I., 1997. Spatial dependence and house price index construction. J. Real. Estate Financ. Econ. 14 (1–2), 203–222.
Soltani, A., Pettit, C.J., Heydari, M., Aghaei, F., 2021. Housing price variations using spatio-temporal data mining techniques. J. Hous. Built Environ. 1–29.
Song, Z., Cao, M., Han, T., Hickman, R., 2019. Public transport accessibility and housing value uplift: Evidence from the Docklands light railway in London. Case Stud. Transp. Policy 7 (3), 607–616.
Thaler, R., 1978. A note on the value of crime control: evidence from the property market. J. Urban Econ. 5 (1), 137–145.
Thibodeau, T.G., 2003. Marking single–family property values to market. Real. Estate Econ. 31 (1), 1–22.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc.: Ser. B (Methodol.) 58 (1), 267–288.
Tita, G.E., Petras, T.L., Greenbaum, R.T., 2006. Crime and residential choice: a neighborhood level analysis of the impact of crime on housing prices. J. Quant. Criminol. 22 (4), 299.
Wang, X., Wen, J., Zhang, Y., Wang, Y., 2014. Real estate price forecasting based on SVM optimized by PSO. Optik 125 (3), 1439–1443.
Wen, H., Xiao, Y., Hui, E.C., Zhang, L., 2018. Education quality, accessibility, and housing price: Does spatial heterogeneity exist in education capitalization? Habitat Int. 78, 68–82.
Wong, T.-T., Yeh, P.-Y., 2019. Reliable accuracy estimates from k-fold cross validation. IEEE Trans. Knowl. Data Eng. 32 (8), 1586–1594.
Worthington, A., Higgs, H., 2013. Macro drivers of Australian housing affordability, 1985-2010. Stud. Econ. Financ.
Wu, C., Hu, W., Zhou, M., Li, S., Jia, Y., 2019. Data-driven regionalization for analyzing the spatiotemporal characteristics of air quality in China. Atmos. Environ. 203, 172–182.
Xu, T., 2017. The relationship between interest rates, income, GDP growth and house prices. Res. Econ. Manag. 2 (1), 30–37.
Xu, T., Zhang, M., Aditjandra, P.T., 2016. The impact of urban rail transit on commercial property value: new evidence from Wuhan, China. Transp. Res. Part A: Policy Pract. 91, 223–235.
Yang, L., Zhou, J., Shyr, O.F., 2019. Does bus accessibility affect property prices? Cities 84, 56–65.
Ye, Y., Xie, H., Fang, J., Jiang, H., Wang, D., 2019. Daily accessed street greenery and housing price: measuring economic performance of human-scale streetscapes via new urban data. Sustainability 11 (6), 1741.
Zhao, Y., Chetty, G., & Tran, D. (2019). Deep Learning with XGBoost for Real Estate Appraisal. 2019 IEEE Symposium Series on Computational Intelligence (SSCI),
Zou, H., Hastie, T., 2005. Regularisation and variable selection via the elastic net. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 67 (2), 301–320.
Zulkifley, N.H., Rahman, S.A., Ubaidullah, N.H., Ibrahim, I., 2020. House price prediction using a machine learning model: a survey of literature. Int. J. Mod. Educ. Comput. Sci. 12 (6).