Cities 131 (2022) 103941

Contents lists available at ScienceDirect

Cities

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

Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms

Ali Soltani [a,c,*], Mohammad Heydari [b], Fatemeh Aghaei [c], Christopher James Pettit [d]

[a] UniSA Business, University of South Australia, Adelaide, 5001, Australia
[b] Department of Urban Planning, Tarbiat Modares University, Tehran, Iran
[c] Department of Urban Planning, Shiraz University, Shiraz, Iran
[d] City Futures Research Centre, University of New South Wales, Sydney, 2052, Australia

ARTICLE INFO

Keywords:
Housing price
Value estimation
Spatio-temporal modelling
Machine learning
Australia

ABSTRACT

Conventional housing price prediction methods rarely consider the spatiotemporal non-stationary problem in a large data volumes. In this study, four machine learning (ML) models are used to explore the impacts of various features – i.e., property attributes and neighborhood quality - on housing price variations at different geographical scales. Using a 32-year (1984–2016) housing price dataset of Metropolitan Adelaide, Australia, this research relies on 428,000 sale transaction records and 38 explanatory variables. It is shown that non-linear tree-based models, such as Decision Tree, have perform better than linear models. In addition, ensemble machine learning techniques, such as Gradient-Boosting and Random Forest, are better at predicting future housing prices. A spatiotemporal lag (ST-lag) variable was added to improve the prediction accuracy of the models. The study demonstrates that ST-lag (or similar spatio-temporal indicator) can be a useful moderator of spatio-temporal effects in ML applications. This paper will serve as a catalyst for future research into the dynamics of the Australian property market, utilizing the benefits of cutting-edge technologies to develop models for business and property valuation at various geographical levels.

1. Introduction

Housing is critical to family well-being, and homeownership is perceived as integral to both the stability of family life and the creation of wealth (Karamujic, 2015). Housing prices reflect the quality of life in urban environments and it is a key element to any sustainable, productive, and resilient city (Ma & Gopal, 2018; Streimikiene, 2015). Spatial and temporal variations are two fundamental factors of housing value changes (Yao & Stewart Fotheringham, 2016). Exploring this variance can help us better grasp how the housing market is changing and is an essential consideration in the formulation and implementation of housing strategy planning and policy setting (Salvati et al., 2019; Sipan et al., 2018). Understanding how the physical environment, as well as the location and physical attributes of a property, impact property value is crucial from a policy viewpoint. Furthermore, it provides understanding of the like spatial and temporal changes of property prices can assist financing mechanisms such as land value capture, which can be used by the government to support transformational urban infrastructure such as new metro lines. As a result, mechanisms like value capture and value sharing may help governments and organisations throughout the world fund infrastructure projects and ease their financial constraints (Diao & Ferreira, 2010).

Housing price variations are studied from many perspectives including in the fields of economics (Belke & Keil, 2018), urban and regional planning (Diao & Ferreira, 2010) and data science (Singh et al., 2020). Excluding broader factors (e.g., macro-economic situation, inflation, recession), housing price variations are influenced by three main categories: (i) Physical characteristics with a focus on dwelling attributes such as building age, building quality, lot area, views, etc. (ii) Location attributes such as accessibility to services including for example - proximity to the central business district (CBD), accessibility to public transport, distance to the closest retail center (iii) Neighborhood and amenity considerations such as crime rates, proximity to good schools and greenspace (Li et al., 2019; Pettit et al., 2020; Wu et al., 2019; Yuan et al., 2020; Zolnik, 2021; Zulkifley et al., 2020).

In housing value research, ordinary least square, OLS multivariate

[*] Corresponding author.
E-mail addresses: ali.soltani@unisa.edu.au (A. Soltani), mohammad.heydari@modares.ac.ir (M. Heydari), f.aghaei@shirazu.ac.ir (F. Aghaei), c.pettit@unsw.edu.au (C.J. Pettit).

https://doi.org/10.1016/j.cities.2022.103941
Received 22 September 2021; Received in revised form 4 August 2022; Accepted 12 August 2022
Available online 25 August 2022

regression is commonly used as a hedonic model (Rosen, 1974). Due to the spatial dependency, however, parameter estimates using traditional hedonic price models might be biased (Anselin, 2013; Chica-Olmo et al., 2019; Yao & Stewart Fotheringham, 2016). To handle spatial heterogeneity, many models applied in the spatial analysis of housing value. A common spatial statistical method for addressing this issue is Geographically Weighted Regression (GWR) (Cao et al., 2019; Fotheringham et al., 2015). The model allows explanatory variables to explain local effects and variations spatially. The GWR model is used in understanding the patterns of many urban phenomena, including real estate market (Harris et al., 2013; Sipan et al., 2018), urban growth (de la Luz Hernández-Flores et al., 2017; Li et al., 2017), urban mobility and accessibility (Qian & Ukkusuri, 2015). Spatial lag and spatial error models are other models that have been used to address spatial dependencies. Moreover, in the case of housing price modelling using historical data, temporal non-stationarity among spatial non-stationarity demonstrates the time-sensitive characteristic in housing price changes (Copiello, 2020; Zhu & Zhang, 2021). Therefore, to account for the temporal dependency in housing price analyses, spatio-temporal models such as geographically and temporally weighted regression (GTWR) have been extended. In comparison to an ordinary GWR model, GTWR appears to be more effective for integrating both spatial and temporal dependencies into the modelling (Zhang et al., 2019).

(Soltani et al., 2021a) examined three prediction models on the real estate transactions data of the Tehran Metropolitan: OLS, GWR, and GTWR. The authors demonstrated that considering spatiotemporal non-stationary is essential in understanding housing price variations and that the GTWR coefficients are more accurate. Another way to addressing spatiotemporal dependencies in data is to consider the spatio-temporal lag variable as an explanatory variable in prediction models. In a semi-parametric spatially weighted regression (S-GWR), Yao and Stewart Fotheringham (2016) introduced a basic spatiotemporal lag variable as an explanatory variable to get better results. Smith and Wu (2009) proposed a spatiotemporal model that took into account both geographical and temporal lag effects, which they used to investigate property price trends in the Philadelphia area of the United States.

Although such models can facilitate an understanding of the changes in space and time, the downside is that the prediction power and precision of these models are normally regarded lower than those which are based on more advanced techniques including machine learning (ML) and deep learning. In recent years, the rise of Big Data, high-performance computing techniques, and advanced ML methods have become a robust prediction approach because it can predict house prices more accurately (Kang et al., 2020; Truong et al., 2020). In many studies that ML models are used for housing price analysis, the spatial and temporal dependencies are neglected in the modelling process. Some ML algorithms can model non-linear effects of explanatory variables on housing price as the dependent variable (Gupta et al., 2021; Yang et al., 2021). The research by Yang et al. (2021) in the Chinese city of Xiamen, indicated that using non-linear ML algorithms has more substantial predictive power than conventional regression methods such as hedonic pricing models. Zhou (2020) investigated how ML models can be used for housing price prediction. The author applied Linear Regression, Lasso Regression, Random Forest and XGBoost, models with 79 features in different categories of variables on the housing dataset in Ames, Iowa, as the case study. Jha et al. (2020) suggested an improved housing price prediction model to predict the housing prices by employing several ML algorithms including Random Forest, CatBoost, XGBoost, and Voting Regressor on the housing datasets of 62,723 records that are acquired from Florida Volusia County Property Appraiser website. In several studies, ML algorithms such as Ridge and Elastic Net regressions (Simlai, 2021), Support Vector Machine (Ho et al., 2021), and Random Forests (Gupta et al., 2021) are applied in housing price prediction research.

While ML models are often used for housing price analysis, the spatially and temporally non-stationarity is neglected in most of the previous research. A study of real estate transactions in New York city, demonstrated that the addition of the "spatial lag" feature to ML models would significantly increase the accuracy and outperform the conventional real estate predictive models (Kiely & Bastian, 2020). In Australia, only a small amount of research has been done on housing price prediction by considering spatiotemporal non-stationarity and using the ML approach (Phan, 2018). On the other hand, the existing studies described the capability of the ML, without deeply investigating the ML models' performance on a real data (Lock et al., 2021). Using 32-year (1984–2016) housing sold price dataset in Metropolitan Adelaide, this paper seeks to analyse the housing price market using ML techniques. Due to the spatiotemporal dependency that commonly exists in housing prices, a spatio-temporal lag variable is applied to the model. In this research, 38 explanatory variables (57 features) in 3 categories are used to predict housing price. Furthermore, to our knowledge, the SAILIS (the South Australian Integrated Land Information System) dataset applied in this study has never been employed in any former housing price prediction research.

The remainder of the paper is organised as follows. The second section covers the fundamental concepts of ML techniques that have been applied in this research and present the modelling framework. Section 3 introduces the case study area, the data collection, and data integration procedure, which are followed by the empirical results and a comparison in Section 4. Finally, in Section 5, the study's conclusion is outlined as well as a discussion of the potential future studies.

2. Methodology

In recent years, ML has risen to prominence as a cutting-edge modelling approach and a subset of artificial intelligence (AI) that enables computers to automatically learn and discover massively complex data in order to detect and comprehend patterns (Ngiam & Khor, 2019). Practically every scientific domain is witnessing the benefits of ML prediction power available for analysing large sets of data (Jha et al., 2020). ML algorithms can be categorized into four classes according to their purpose: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In a supervised learning model, after dividing the train and test dataset, the algorithm can use a labelled dataset to train the model and evaluate its accuracy on test data. Unlike supervised learning, an unsupervised machine learning model doesn't require labelled data and the algorithm tries to discover patterns by identifying features based on their shared characteristics. Based on these definitions, to solve the housing price prediction problem, supervised ML regression models are mostly commonly applied. In this research, we choose four supervised ML algorithms to illustrate how ML can be applied to predict housing prices with the combination of different types of attributes (dwelling variables, accessibility variables, and neighborhood variables); moreover, the effect of spatiotemporal lag on the accuracy of used models is investigated. These four regression models include Linear Regression, Decision Tree, Random Forest, and Gradient-Boosted Tree. The Grid Search Cross-Validation method has been used for hyper-parameter tuning of these ML models. The accuracy of models is evaluated using the adjusted-R2, root mean square error (RMSE), and mean absolute error (MAE) on the split train and test dataset. Fig. 1 represents the flowchart of research process.

In this research, the first step is the use of the ArcGIS desktop application for data integration and variable generation; secondly, the PySpark framework is used for data-preprocessing and data processing using PySpark SQL module and Spark's MLlibrary (MLlib) respectively.

2.1. Linear regression

The linear regression is regarded as one of the most popular ML algorithms. To fit the model, the linear regression model presumes a linear connection between the dependent and explanatory variables. When there is only one independent variable the method is referred as simple

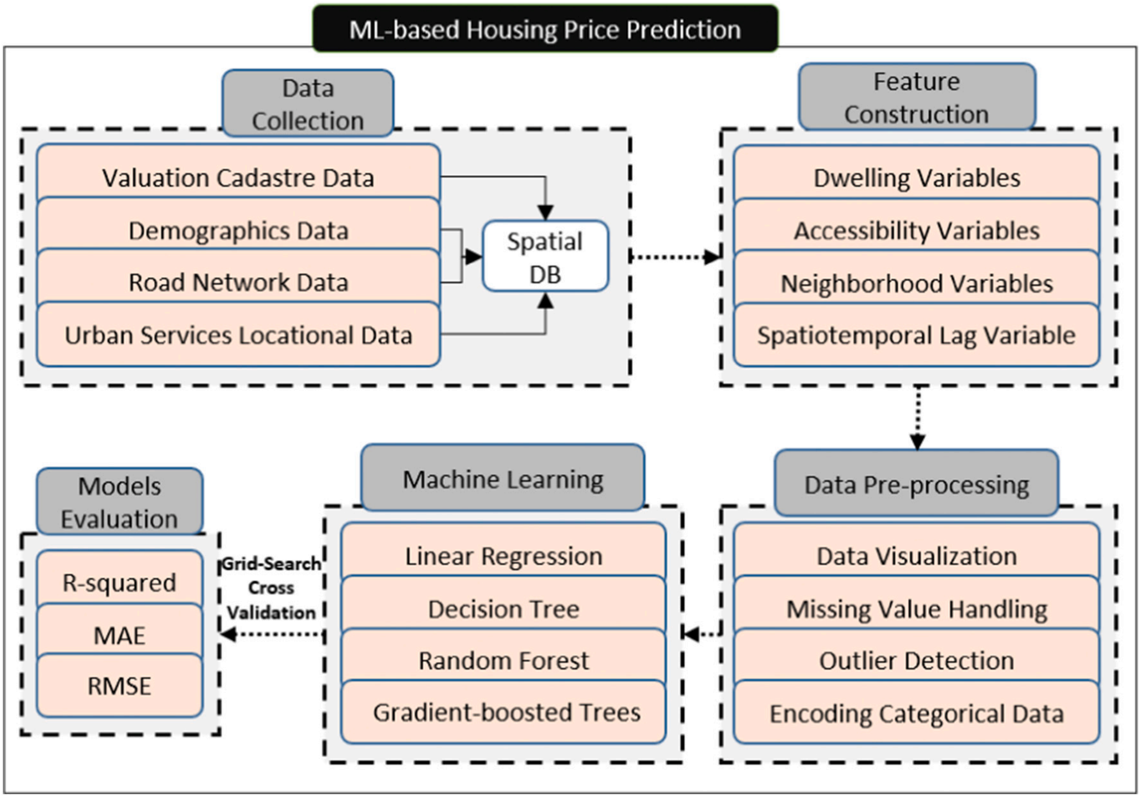


Fig. 1. The flowchart of research process.

linear regression. When there is more than one independent variable, it refers as multiple linear regression.

Given a dataset $\{\boldsymbol{y_i}.\boldsymbol{x_{i1}}.\ \ldots.\boldsymbol{x_{ip}}\}_{i=1}^{\boldsymbol{n}}$ of $\boldsymbol{n}$ rows, the linear regression equation considers a linear association between the dependent variable y and the ***p-vector*** of explanatory variables x.

The error term, ϵ, remainder term, residual or disturbance which is an unobserved random variable that adds "noise" to the fit line and compensates for a lack of complete goodness of fit. Thus the linear regression formula is as below (Zhou, 2020):

$$\mathrm{y_i} = \beta_0 + \beta_1 \mathrm{X_{i1}} + \beta_2 \mathrm{X_{i2}} + \ldots + \beta_k \mathrm{X_{ik}} + \varepsilon_i$$

where, $\boldsymbol{\beta_0}$ represents the constant value, $\boldsymbol{X_k}$ is the ***k-th*** independent variable, $\boldsymbol{\beta_k}$ is the coefficient of the ***k-th*** independent variable and $\boldsymbol{\varepsilon}$ is the residual value. Because of ease to determine and basic assumption of linear relationships between variables, the linear regression has been the first form of regression model to be thoroughly investigated and widely used in housing value research practice.

2.2. Decision Tree

Decision Tree (DT) is a non-parametric supervised ML algorithm that is applied to classification and regression issues. DT is applied to make predictions by training simple decision rules inferred from the data characteristics. DT generates a set of if-then-else decision rules by Learning from data. The complexity of the model is measured by the depth of tree and generally, the deeper tree has the more complex decision rules (Xu, 2015). DT is popular due to its ease of interpretation, ability to handle categorical features, lack of feature scaling, and ability to capture non-linear relationships between dependent and independent variables.

The DT is a greedy method that divides the feature space into binary divisions recursively. The tree predicts the same label for each bottommost (leaf) division. To maximize the information gain at a tree node, each partition is selected greedily by picking the best split from a collection of potential splits. In other words, for each tree node, the split is picked from the set $\boldsymbol{argmax_s IG(D.s)}$ where $\boldsymbol{IG(D.s)}$ represents the information gain when a split $\boldsymbol{s}$ is applied to a dataset $\boldsymbol{D}$ (Meng et al., 2016). The label homogeneity at the node is measured by the node impurity. The following is an impurity measure for a decision tree regressor:

$$\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_i}-\mu)^2$$

where; $\boldsymbol{y_i}$ is label for a row of data, $\boldsymbol{N}$ is the count of data rows and $\boldsymbol{\mu}$ is the average given by $\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y_i})$.

The decrease in entropy or surprise achieved by changing a dataset is known as information gain, and it is frequently utilised in the training of decision trees.

The difference between the parent node impurity and the weighted sum of the two child node impurities is used to calculate the information gain, which tells us how essential a particular feature vector property is. Assuming that a split $\boldsymbol{s}$ divides the dataset $\boldsymbol{D}$ of size $\boldsymbol{D}$ into two datasets $\boldsymbol{D_{left}}$ and $\boldsymbol{D_{right}}$ of sizes $\boldsymbol{N_{left}}$ and $\boldsymbol{N_{right}}$, respectively, the information gain is:

$$\mathrm{IG(D,s)} = \mathrm{Impurity(D)} - \frac{\mathrm{N_{left}}}{\mathrm{N}}\mathrm{Impurity}\left(\mathrm{D_{left}}\right) - \frac{\mathrm{N_{right}}}{\mathrm{N}}\mathrm{Impurity}\left(\mathrm{D_{right}}\right)$$

One downside of DT is that it is prone to overfitting, which means that projected outcomes may be influenced incorrectly by atypical examples. However, this problem may be overcome using ensemble techniques. An ensemble technique is a learning process that generates a model from a collection of other models. In this research two decision tree based ensemble algorithms are applied that include: Random Forest and Gradient-Boosted Trees.

2.3. Random Forest

Ensemble learning methods are intended to increase accuracy and robustness over a single estimator by incorporating the estimated results of a set of base estimators. Random Forest proposed by Breiman (2001) is a supervised machine learning algorithm which uses the ensemble learning method for classification and regression. It iteratively runs a number of decision trees that trained across randomly generated subsets of data, and combines them into a single model to make more precise prediction while declining the risk of overfitting. Random forests as highly scalable and parallelisable ensemble method, in regression problem, train a set of decision trees separately and then employing averaging to generate predictions. In other word, the label is predicted to be the mean of the tree predictions.

The method introduces randomisation into the training process, making each decision tree unique. The variance of the predictions is reduced when the predictions from each tree are combined, which improves the performance on test data. Random Forest algorithm like Decision Tree is able to capture non-linear relationships between dependent and independent variables. This is extremely useful for capturing non-linear relationships in housing value analysis because variations in space and time increase the likelihood of non-linear relationships (Kiely & Bastian, 2020).

2.4. Gradient-Boosted Tree

Gradient-Boosted Trees (GBTs) is an ensemble ML method in the form of an ensemble of weak prediction models based on decision trees. GBTs train decision trees iteratively in order to minimize a loss function. Gradient boosting is a method for training a series of DTs repeatedly. on each iteration, this algorithm uses the current ensemble to predict the label of each training row data, and then compares the prediction to the real label. The dataset has been re-labelled to emphasize on training cases with poor predictions. As a result, the decision tree assists in correcting prior errors in the following iteration. A loss function defines the specific technique for re-labeling values. GBTs lower the loss function on the training data with each iteration. In this research Squared Error loss function is used with the following equations:

$$\text{Squared Error} = \sum\nolimits_{i=1}^{N} (y_i - F(x_i))^2$$

where; $\boldsymbol{N}$ is the number of row data, $\boldsymbol{y_i}$ is label of row data $\boldsymbol{i}$. $\boldsymbol{x_i}$ is features of row data $\boldsymbol{i}$. $\boldsymbol{F(x_i)}$ is model's predicted label for row data $\boldsymbol{i}$.

In comparison with Random Forest, GBTs train one tree at a time, taking longer than random forests to complete. Random Forests may simultaneously train several trees. Overfitting is less likely with Random Forests. Overfitting is less likely when more trees are trained in a Random Forest, but it is more likely when more trees are trained with GBTs (Spark, 2021).

2.5. Hyperparameter tuning

Each of the ML algorithms that is used in this research has a set of parameters which are referred as hyperparameters. Generally, hyperparameters are chosen during the validation process. Cross-validation is regarded as an effective method for identifying the best parameters.

The initial stage in the cross-validation procedure is to divide the dataset into folds, which will be utilised as separate training and test datasets. For example, if k = 3 folds, data will be split into three dataset pairs, with two-third of the data used for training and one-third for testing during the modelling phase. Finally, to evaluate the models, computes the average evaluation metric for the three models created by fitting the estimator method to the three dataset pairings.

There are two methods for selecting parameters: grid search and random search. In fact, grid search is a way of exhaustively scanning a manually defined section of the hyperparameter space of a given algorithm. Random search, on the other hand, uses a probability distribution to choose a value for each hyperparameter. In this research, Grid-search Cross-Validation is used to hyperparameter tuning for the purpose of maximize the accuracy of ML algorithms. Hyperparameters tuning is helpful to better understanding and evaluation between used ML algorithms (Xu, 2015). Table 1 summarizes the needed parameters that must be optimized for used ML algorithms in this paper.

2.6. Model Evaluation

Model Evaluation is a critical step in ML process and has been used with different measures. It specifies how accurate the fitted models perform in predicting process. Three types of accuracy measures are applied: R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These are the formulae for these measures:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} \| \hat{y}_i - y_i \|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

where n refers to the number of data rows, $\hat{y}_i$ is the predicted label and y_i is the actual label, $\bar{y}_i$ is mean value of y.

2.7. Spatiotemporal lag variable

For each of the sold properties as the dependent variable, a spatiotemporal lag variable is calculated using a customized Inverse Distance Weighting (IDW) method and considering spatiotemporal neighbor selection. To do this, for each sold property point, all points within a temporal search radius of 6 months and a spatial search radius were selected as neighbors, and then the spatiotemporal lag was calculated using the IDW method (Fig. 2). The rational for choosing 400 m is that this is the likely convenient walking distance which is used in defining the domain of a residential neighborhood based on Clarence Perry (1929) idea (Lloyd Lawhon, 2009). With an annual average home price rise of 5 % in South Australia, the six-month period was deemed a suitable time to see a considerable increase in property prices. The structure of the customized IDW method that is used in this research can be expressed as:

Where, ST_i refers to spatiotemporal lag for property value i; $sd_{[i.\ j]}$ refers to spatial distance between i. j; r_s refers to spatial search radius; $td_{[i.\ j]}$ refers to temporal distance between i. j; r_t refers to temporal search radius; and hp_j refers to property value of j (the unit price of a property per square meter).

Fig. 3 shows the descriptive statistics and distribution of the calculated ST-Lag variable against the dependent variable (price/meter). As it shows there are approximately linear relationships between these variables. Also, distributions of both variables are shown in the left chart that indicating distribution similarity between them.

Table 1
Hyperparameters for used ML algorithms.

ML algorithms	Hyperparameters
Decision Tree	Minimum sample leaf; minimum samples split; maximum features; maximum depth
Random Forest	Number of estimators; minimum samples split; minimum sample leaf; maximum features; maximum depth
Gradient-Boosted Trees	Learning rates; the number of estimators; minimum samples split; minimum sample leaf; maximum features; maximum depth

$$ST_i = \frac{\sum_{j|sd_{[i.j]}<r_s;td_{[i.j]}<r_t}\frac{1}{sd_{[i.j]}} * hp_j}{\sum\frac{1}{sd_{[i.j]}}}$$

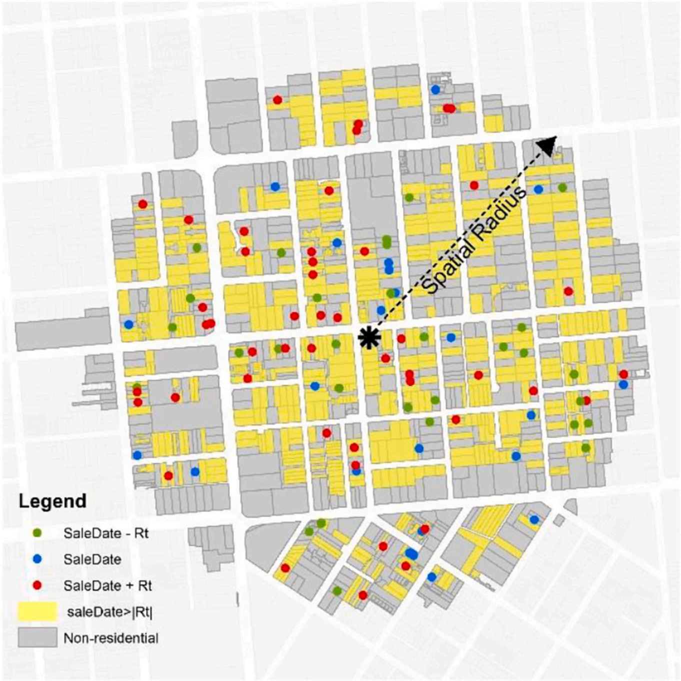
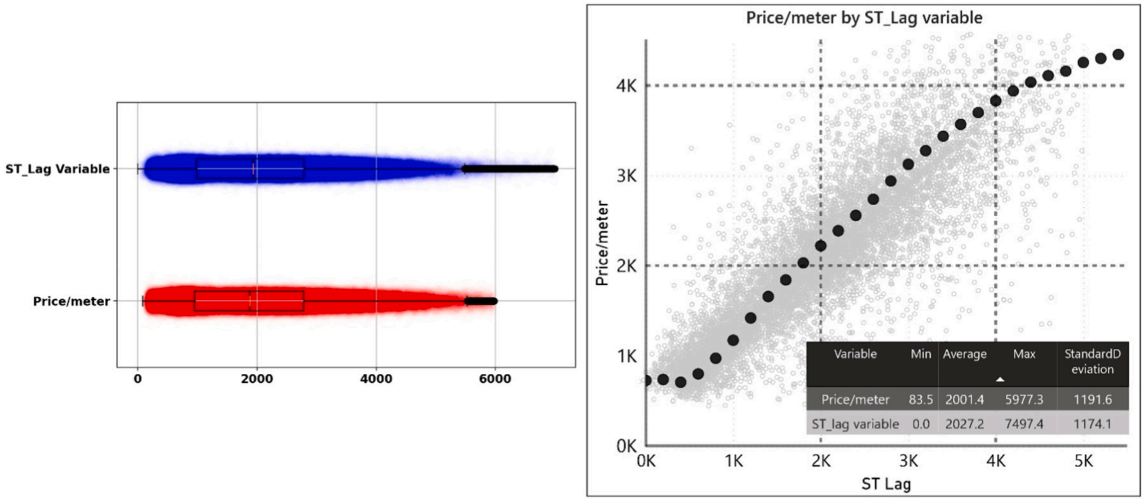


Fig. 2. The spatial and temporal search radius.



Variable	Min	Average	Max	StandardDeviation
Price/meter	83.5	2001.4	5977.3	1191.6
ST_lag variable	0.0	2027.2	7497.4	1174.1

Fig. 3. Descriptive statistcs of ST-Lag variable.

3. Data collection

3.1. Case study area

In this paper, the real estate data comes from the South Australian Integrated Land Information System (SAILIS) for the period of 32 years (1984–2016). Other required data included socio-economic status data are downloaded from the Australian Bureau of Statistics (ABS) and South Australian Government Data Directory (Data.SA). Adelaide with a land area of 3259.8 km^2 is the capital of South Australia, and the fifth-most populous city of Australia (>1.4 million) (Nguyen et al., 2018; Soltani et al., 2022a). Adelaide is an important destination for immigrants. The high levels of immigrations affects the housing price growth and makes housing market less affordable in Adelaide (Moallemi & Melser, 2020). The cost of housing was discovered to be an important determinant in homeowner's travel behaviour (Soltani et al., 2021b; Soltani et al., 2022b). The average housing price in Adelaide varies in both spatial and temporal dimensions. The growth rate of housing price per square metre for the research period was 390 % (Fig. 3-c). Also, spatial variations of property price over research time interval is shown in Fig. 3. We have selected all of 428,000 sold cases from all residential property types in the metropolitan area and in research time period. The spatial distribution of selected cases in suburbs and the temporal distribution of them are presented in Fig. 3-a.

Fig. 4 shows the fluctuations over time (between 2000 and 2016). Fig. 4b shows that the rise in housing values has mainly occurred on the beach, in some parts of the south, and in the central area of the CBD as before. In this research, based on the literature review and data accessibility, 38 explanatory variables in three categories are used to predict housing price using a ML procedure. Also, the spatiotemporal lag variable is added as an explanatory variable to considering spatiotemporal non-stationarity. A descriptive summary of the used variables is

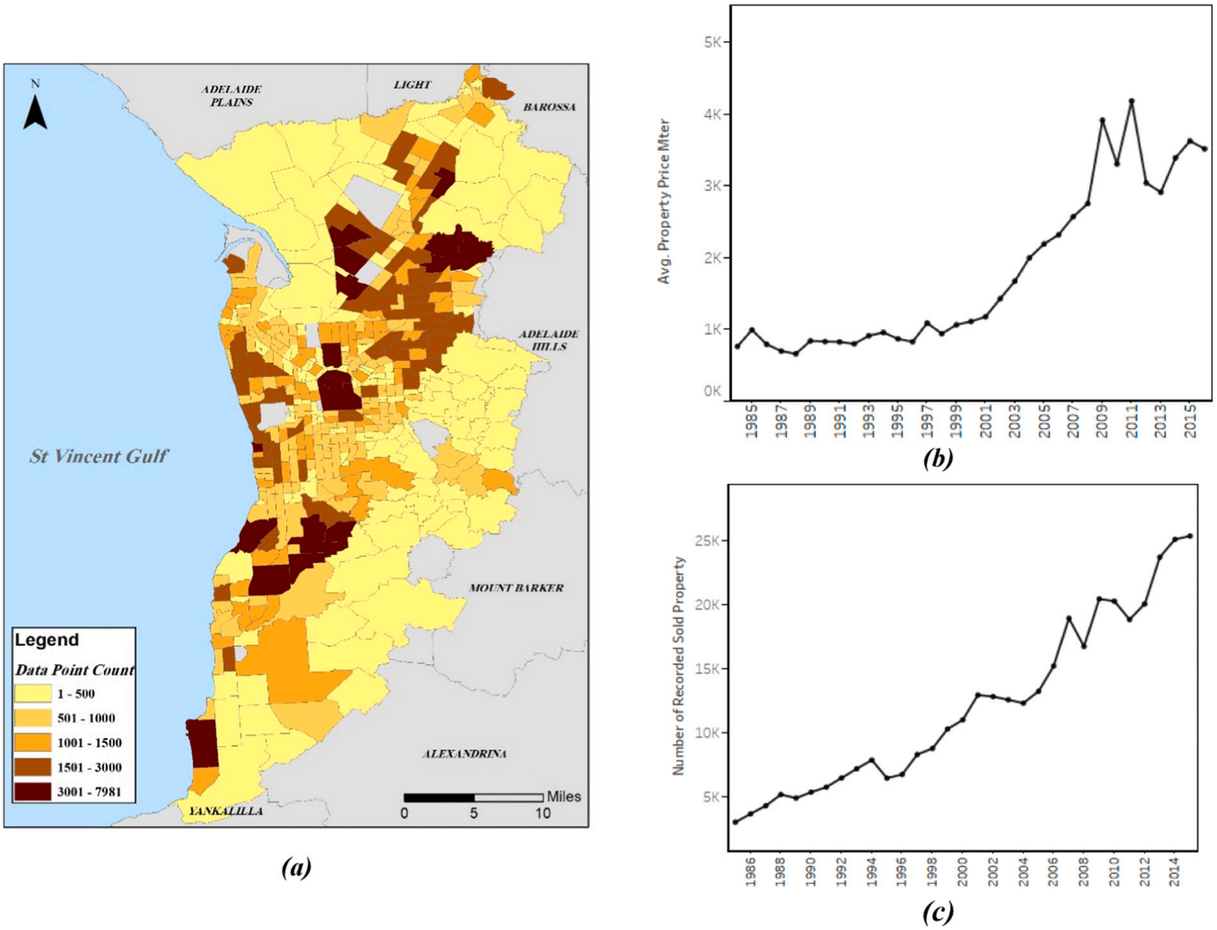


Fig. 4. a) Case Study area and spatial distribution of sold residential properties; b) annual frequency of sold residential properties; c) annual average of property price.

presented in Table 2.

4. Results

4.1. Data pre-processing

Before applying the ML algorithms, data preprocessing is an integral step to enhance data reliability; because the quality of data directly affects the accuracy of learning. The explanatory variables in regression equations are called features in ML terminology and feature engineering refers to obtaining new features from old features to improve the performance of ML algorithms. Data preprocessing and feature engineering in this research are included as:

I. Remove the rows with missing value in sale price and sale date

II. Remove the rows with missing value in the dwelling variables [x1:x15]

III. Encoding categorical variables to numerical variables based on cardinality measure. Ordinal encoding technique is applied to *Building quality* variable. One-Hot Encoding technique is used to encoding the variables with low cardinality level (*Sewerage availability*, *water availability*, *ownership type of dwelling*, *dwelling type*, *roof material type*) while Frequency encoding technique is used to encoding variables with high cardinality level (*Wall material type*, *Building style*). After variable encoding we have had 64 features to use in ML models.

IV. Outlier detection aid in making the training model more stable. In this research, Inter-Quartile Range (IQR) method is used to detect outlier data rows on columns. Based on this IQR method, an outlier $\boldsymbol{x}$ can be detected if:

$$x < Q_1 - 1.5 \times \text{IQR OR } Q_3 + 1.5 \times \text{IQR} < x$$

a. $Q_1 = 25$th percentiles; $Q_3 = 75$th percentiles; $IQR = Q_3 - Q_1$

V. Remove features that might lead to multicollinearity. In order to recognise features that cause multicollinearity, a HeatMap is used to plot correlation between all 57 features as shown in Fig. 5. As results, 11 features (x9-1, x10-1, x11-1, x12-2, x13-9, x17, x24, x26, x30, x34, x35) are dropped from further steps.

VI. Finally, Min-Max scaling method is used to normalize data which feature values are rescaled to a range between 0 and 1.

After data pre-processing, the final dataset contained 352,024 data points with 46 features. In total, 38 variables were as input to the data preprocessing and 46 features were as output to data processing (Fig. 6).

4.2. Linear regression results

First, multiple linear regression is applied to recognise the significant features which affect the housing price assuming a linear relationship between features and target variable. Before training the model, the final dataset has randomly been split into two sets for the modelling process. The training set and test set have been contained 70 % and 30 %, respectively. To show the effect of spatiotemporal lag variable on the model performance, the model is trained once without spatiotemporal lag variable and again with incorporating this feature to the model. After

Table 2
The description of the variables.

Characteristic Type	Variables	Description	Type (unit)	Data source
Dwelling variables	x1	The age of construction dwelling when property is sold	int (year)	SAILIS
	x2	Building quality	str (ordinal grades (1–9))	SAILIS
	x3	Number of ensuite room	int (count)	SAILIS
	x4	Dwelling area	Double (square meter)	SAILIS
	x5	Floor Level	int (count)	SAILIS
	x6	Number of bedroom	int (count)	SAILIS
	x7	Number of main room	int (count)	SAILIS
	x8	Number of stories in the building	int (count)	SAILIS
	x9	Sewerage availability	Boolean (Y/N)	SAILIS
	x10	Water availability	Boolean (Y/N)	SAILIS
	x11	Ownership type of dwelling	str (4 categorical type)	SAILIS
	x12	Dwelling type	str (7 categorical type)	SAILIS
	x13	Roof material type	str (9 categorical type)	SAILIS
	x14	Wall material type	str (14 categorical type)	SAILIS
	x15	Building style	str (49 categorical type)	SAILIS
Accessibility variables	x16	Adjacent Street type	str (56 categorical type)	data.sa.gov.au
	x17	Euclidian[a] distance to O-bahn subway	Double (metric distance)	data.sa.gov.au
	x18	Network distance to nearest airport	Double (metric distance)	SAILIS
	x19	Euclidian distance to beach	Double (metric distance)	Authors
	x20	Euclidian distance to CBD	Double (metric distance)	SAILIS
	x21	Network distance to nearest Primary school	Double (metric distance)	SAILIS
	x22	Distance to nearest Roadside Significant Sites	Double (metric distance)	data.sa.gov.au
	x23	Network distance to nearest secondary school	Double (metric distance)	SAILIS
	x24	Network distance to nearest university	Double (metric distance)	SAILIS
	x25	Network distance to nearest train line stops	Double (metric distance)	ABS
	x26	Network distance to nearest tramline stops	Double (metric distance)	ABS
	x27	Network distance to nearest hospital	Double (metric distance)	SAILIS
	x28	Network distance to nearest public space	Double (metric distance)	SAILIS
	x29	Network distance to nearest main shopping center	Double (metric distance)	SAILIS
	x30	Network distance to nearest marina berth	Double (metric distance)	SAILIS
	x31	Network distance to nearest entertainment center	Double (metric distance)	SAILIS
Neighborhood variables	x32	The Index of Education and Occupation (IEO)	int (ordinal grades (1–10))	ABS (SEIFA[c])
	x33	The Index of Economic Resources (IER)		
	x34	The Index of Relative Socio-economic Advantage and Disadvantage (IRSAD)		
	x35	The Index of Relative Socio-economic Disadvantage (IRSD)		
	x36	Population per area of SA1[b] geographical unit that the sold property is located	Double (people/hectare)	ABS
	x37	Population Change in the research period	Double (percent)	ABS
Spatiotemporal Lag	x38	Spatiotemporal Lag	Double	Authors

[a] The reason for using Euclidean distance instead of the Network is that the destination was some kind of polygon instead of a point or a line.
[b] SA1: Statistical Areas Level 1.
[c] SEIFA: Socio-Economic Indexes for Areas.

training the model on training dataset, the test dataset is used for model evaluation. The Adjusted-R2, MAE and RMSE of both the training and the test sets are presented in the.

Table 3. The results show that the model with incorporation of spatiotemporal lag (ST-lag) variable improve of the model's performance; in other words, the model that incorporate ST-lag had the lower RMSE and MAE but higher R^2.

The results are detailed in Table 4. Based on the t-statistics and *p*-value, the coefficients of 39 features are statistically significant in predicting housing price at the 99 % confidence level. According to the coefficient value for each feature, we can interpret how a feature affects housing prices. For example, the *building quality*, *number of ensuite room*, *network distance to nearest train line stops*, *the Index of Education and Occupation (IEO)*, *the Index of Economic Resources (IER)* influence housing price positively, whereas *dwelling area*, *Euclidian distance to nearest beach and Euclidian distance to CBD* influence the housing price negatively.

Before we can explain the impact of important factors on home prices, we must first meet the linear regression assumptions. To fulfil findings, four fundamental assumptions must be met: i) nil or little multicollinearity; ii) no auto-correlation; iii) normal distribution of residuals; and iv) homoscedasticity.

The Variance Inflation Factor (VIF) score for each feature implies that there is no multicollinearity between features (VIF 5) (Table 4). Second, the Durbin-Watson test was employed to analyse the model's residual autocorrelation. The Durbin-Watson test score for residuals was 1.97 (values between 1.5 and 2.5 are acceptable). Third, a histogram chart was used to determine if the residual distribution was normal or

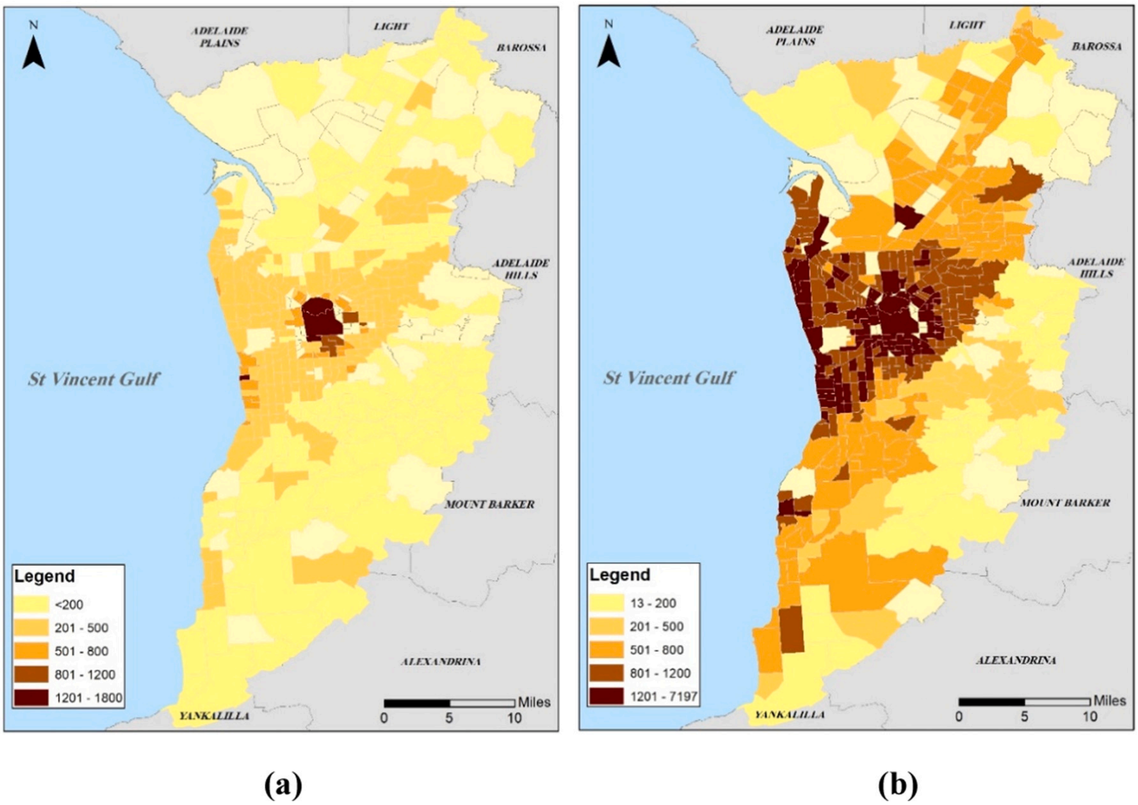


Fig. 5. Spatial variations of property price over research time interval: a) year 2000; b) year 2016.

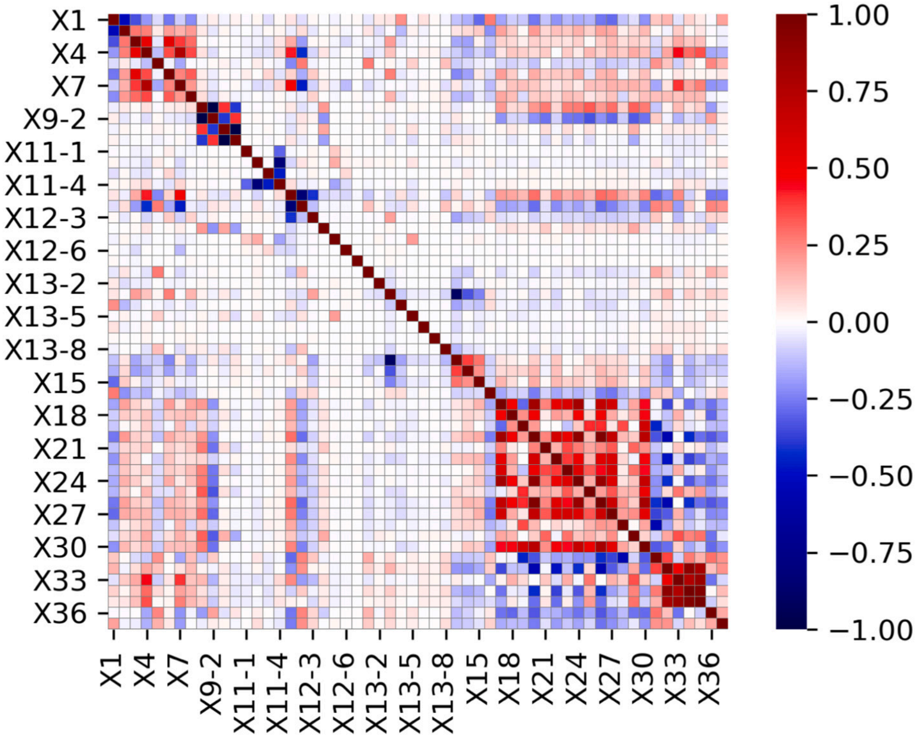


Fig. 6. Correlation matrix between features.

not, and we compared it to a normal distribution curve with a bell shape. In addition, no criteria or significance tests for skewness and kurtosis should be used in big sample data (Field, 2013; Ghasemi & Zahediasl, 2012; Micceri, 1989). The results of producing a histogram chart of the frequency of standardised residuals (test set) reveal that the residual distribution is normal, with a mean and median value of zero. We sought to discover heteroscedasticity (the violation of homoscedasticity) by displaying residuals against predicted values using a technique known

Table 3
Prediction accuracy of multiple linear regression.

Model			Adj-R^2	MAE	RMSE
Multiple linear regression	−ST_lag	Train set	0.462	0.114	0.149
		Test set	0.451	0.121	0.152
	+ST_lag	Train set	0.664	0.087	0.114
		Test set	0.650	0.090	0.129

as the "Wandering Schematic Plot." In this situation, heteroscedasticity does not occur, and the variance of residuals is the same across all predicted values.

Following satisfaction with the linear regression findings, in order to demonstrate how relevant elements influence property prices, the scatterplots of each variable versus price/metre as the dependent variable were plotted (Figs. 7–9).

Because various relationship trends are non-linear, we employed non-linear linear machine learning regression models to produce findings. For example, while the general trend line of "building area" vs "Price/meter" is straight with a positive slope, we can detect two local trends as a non-linear association. A declining trend (negative effect) for properties with less than thirty years of age and an upward trend (positive effect) for properties with more than thirty years of age.

Also, for categorical variables, we utilised dispersed plots of average housing price per each different value of variable and showed the findings to show whether particular categorical factors had significant influence on the dependent variable or not. For example, using "building style" as a major variable, the average housing price for the "high quality conventional" type is 25 % higher than the "conventional" type and 40 % higher than the "SAHT conventional type," as shown in Fig. 7. Significant factors such as "Dwelling Area," "Number of Main Rooms," and "Number of Bedrooms" have a negative influence on home prices; in other words, as Dwelling Area increases, the price per square metre decreases. The building quality has a significant positive link with property prices.

Aside from dwelling characteristics, two more types of variables (neighborhood and accessibility variables) are utilised to predict home prices. As seen in Fig. 10, socioeconomic factors such as "IRSD," "IEO," and "IRSAD" have a beneficial influence on home prices. Housing prices rise in tandem with the increase in "population density" among communities. Prices are affected by the type of surrounding street for each property. The average house price adjacent to the "RD type" of the street is 14 % more than the average housing price adjacent to the "ST type." When it comes to accessibility factors, "distance to the CBD" has a negative influence on house value. This suggests that going farther from the CBD lowers property price; but, at distances >20,000 km, the influence of this variable is considerably decreased.

In general, the distance from the beach and home prices have an inverse connection. This link, however, is rather nonlinear, as seen in the diagram. Distance from public areas has a detrimental impact on house costs as well. In other words, individuals prefer to buy closer to home than in public places. The proximity of elementary schools has a non-linear influence on house values. In other words, at distances between (zero to 500 m, distance from schools has a positive influence on pricing, and distances >500 m, this variable has a negative effect on prices), maybe due to noise and congestion that can develop up to 500 m and the cost of access to school might be affected and raised as the distance increases. The distance from hospital (distance to hospital) is also inversely correlated to housing prices in general. However, due to the pollution and congestion created by the hospital, this connection is direct at distances <200 m, while it is typically inverse at distances >200 m.

4.3. Non-linear regression results

After specifying significant variables using multiple linear regression model, non-linear ML algorithms have been used to improve prediction accuracy. Three tree-based algorithms are applied: Decision Tree and two ensemble methods: Random Forest and Gradient-Boosted Tree. These methods have hyperparameter that must be tuned when applying for prediction. We used 5-fold grid search cross-validation to hyperparameter tuning process. Table 5 presents the evaluation of these algorithms on test dataset, together with the performance metrics, R2, RMSE and MAE.

First, using the DT model, measures are estimated to be 0.797, 0.094, and 0.064, respectively, implying a reasonable good fit.

The results show that the decision tree model as a non-linear ML method improves the learning accuracy and leads to significantly better performance than Multiple Linear Regression models. In most circumstances, the RF model outperforms decision trees. Because, decision

Table 4
The results of linear regression model.

Variable	Coefficient	P-value	VIF	Variable	Coefficient	P-value	VIF
Intercept	−0.008326	0.001249*	–	X13-5	0.004499	0.337026	1.074588
X1	0.291772	0.000000*	2.077386	X13-6	−0.00731	0.07663	1.010722
X2	0.187791	0.000000*	1.581032	X13-7	0.000543	0.944918	1.003947
X3	0.270876	0.000000*	1.915588	X13-8	0.020064	0.000000*	1.074234
X4	−0.455667	0.000000*	3.75927	X14	−0.016975	0.000000*	1.329423
X5	0.250571	0.000000*	1.246935	X15	0.016118	0.000000*	1.247556
X6	0.211067	0.000000*	1.686434	X16	−0.044254	0.000000*	1.199214
X7	−0.390976	0.000000*	2.904713	X18	−0.006989	0.000010*	1.756519
X8	0.091942	0.000000*	1.38388	X19	−0.013689	0.000000*	2.232602
X9-2	−0.008477	0.000000*	1.612007	X20	−0.027328	0.000000*	5.321059
X10-2	0.021167	0.000000*	1.250048	X21	0.034235	0.000000*	1.293516
X11-2	0.018842	0.000001*	1.019546	X22	−0.03412	0.000000*	2.015001
X11-3	0.032071	0.000000*	1.04414	X23	−0.002469	0.436115	2.240052
X11-4	0.039333	0.000000*	1.039539	X25	0.042063	0.000000*	2.019505
X12-1	−0.005835	0.000000*	2.298797	X27	0.014279	0.000004*	3.355123
X12-3	0.008516	0.000000*	1.382405	X28	0.029082	0.000000*	1.593862
X12-4	0.036667	0.000000*	1.10889	X29	−0.007502	0.015926*	1.470489
X12-5	0.037568	0.000000*	1.098187	X31	0.023286	0.000000*	2.401997
X12-6	−0.001157	0.822685	1.07009	x32	0.062963	0.000000*	3.829678
X12-7	−0.00289	0.852759	1.003528	x33	0.042262	0.000000*	2.54601
X13-1	0.058478	0.000000*	1.152398	X36	0.038962	0.000000*	1.466323
X13-2	−0.001321	0.60048	1.018707	x37	−0.027763	0.000000*	1.08017
X13-3	−0.007918	0.000000*	1.422489	x38	3.600779	0.000000*	1.237739
X13-4	−0.023672	0.000000*	1.107155				

* Significant at 99 % level ($p < 0.01$).

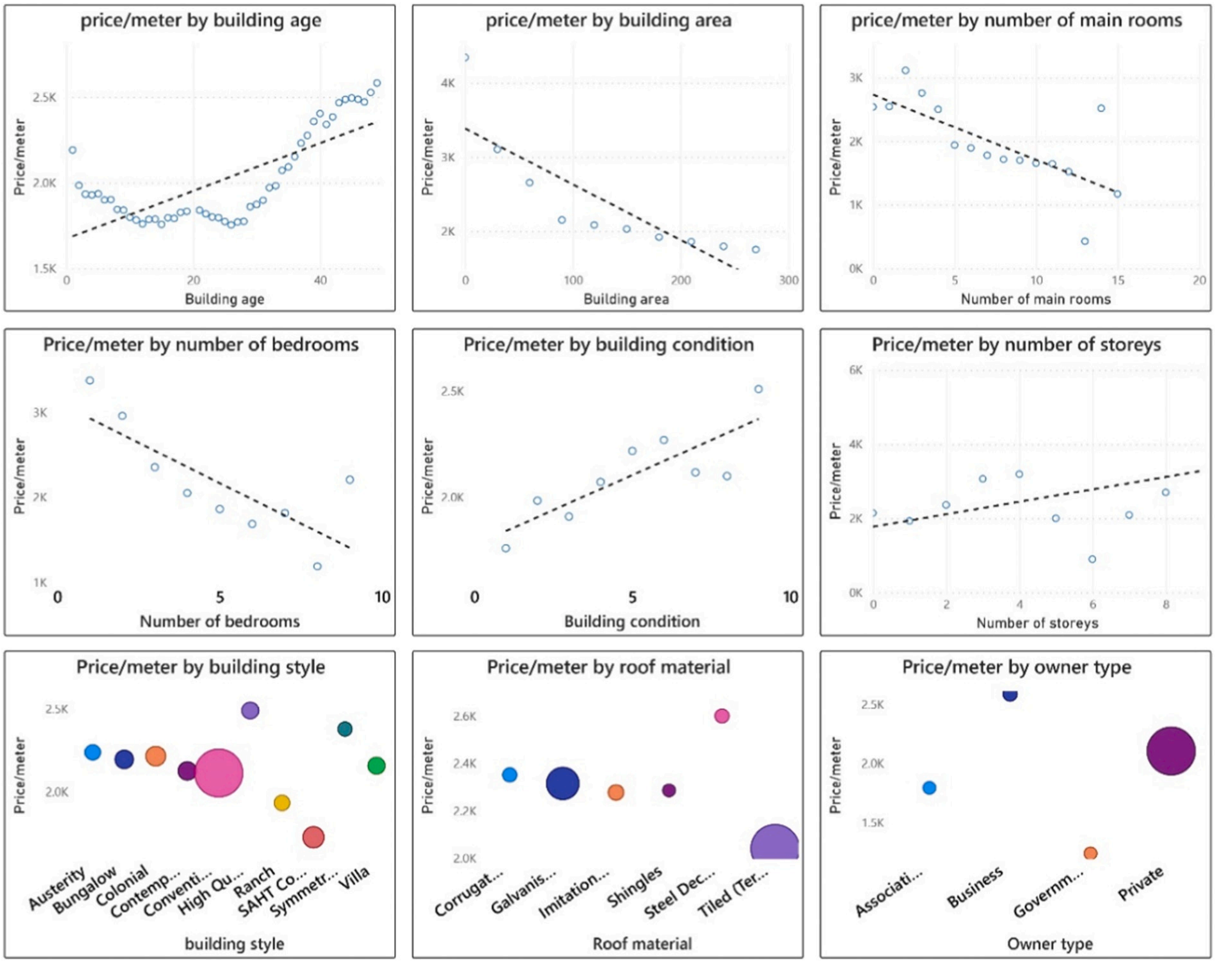


Fig. 7. Scatter plot of dwelling variables against dependent variables.

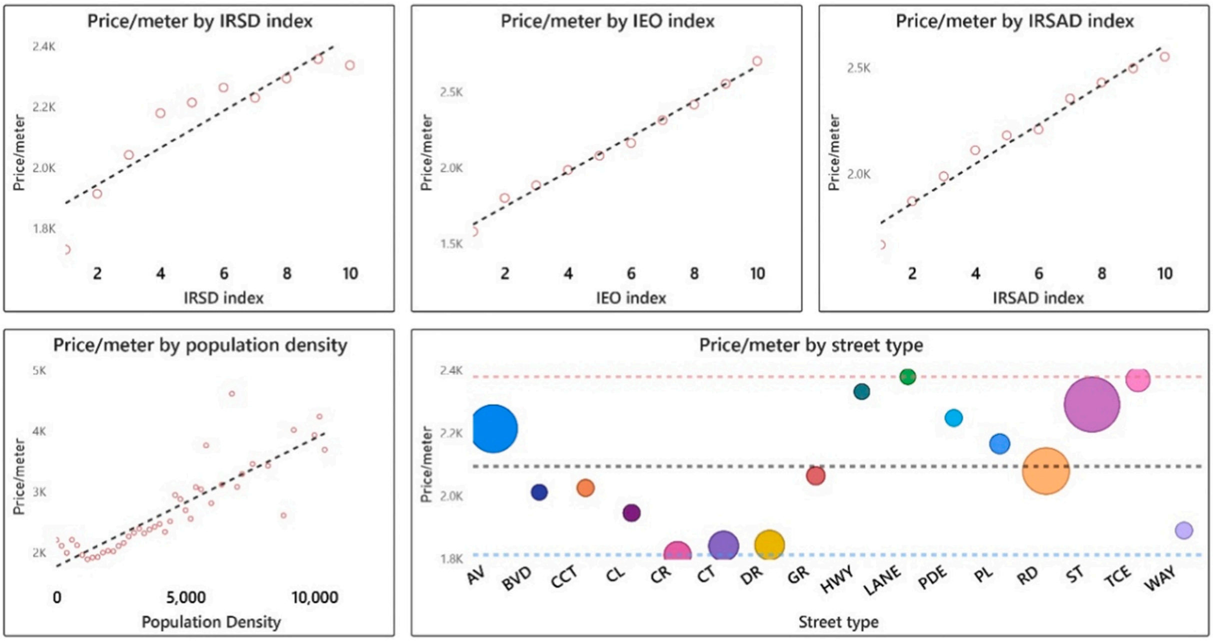


Fig. 8. Scatter plot of neighborhood variables against dependent variables.

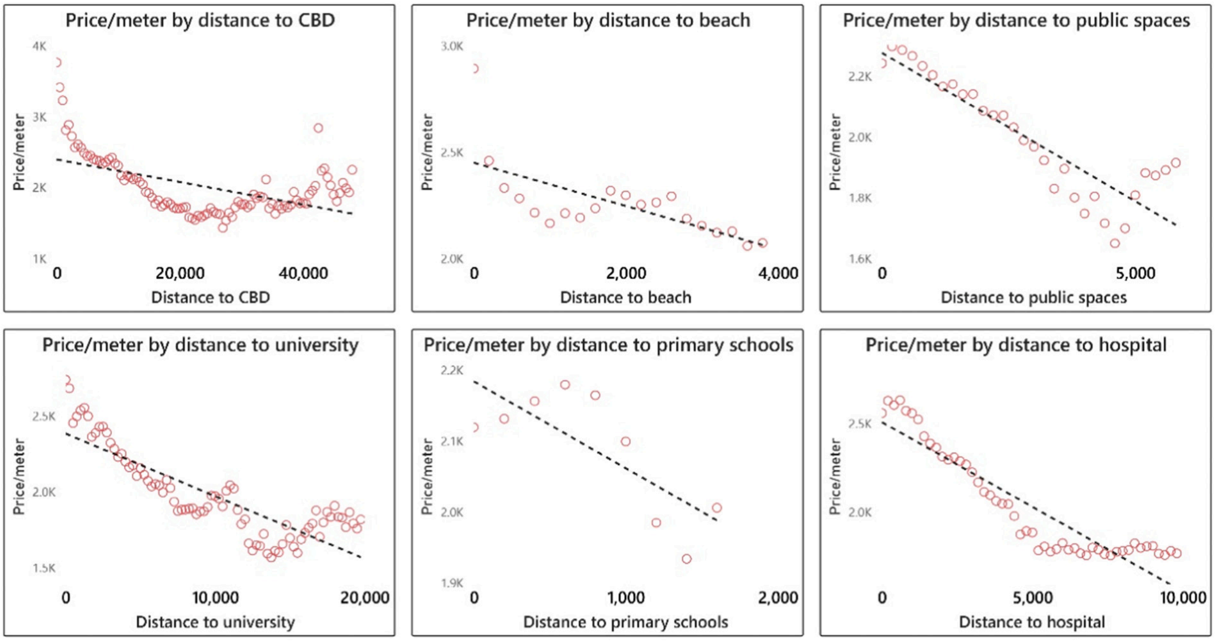


Fig. 9. Scatter plot of accessibility variables against dependent variables.

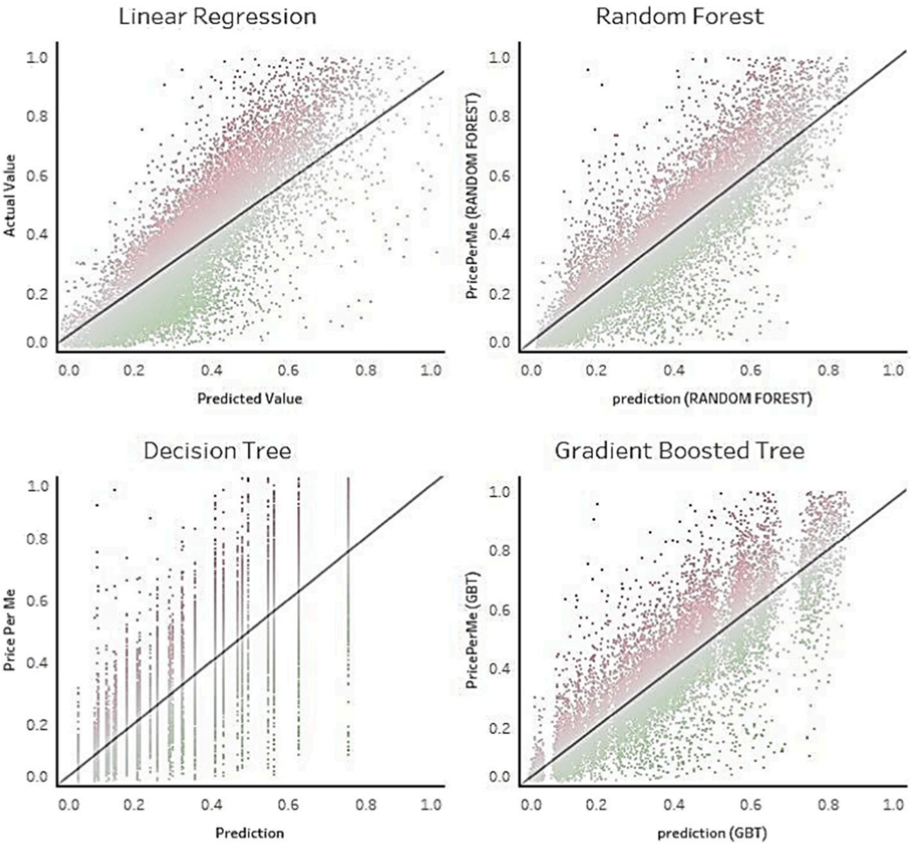


Fig. 10. Comparison of Models Estimation (scatter plots between actual value and predicted value).

trees are more prone to overfitting. Random forests are favored because they offer higher generalization (Xu, 2015). Using RF model through 5-fold cross validation and grid search on hyperparameters, our model provides Adjusted R^2 of 0.875 for the test set. The RMSE and MAE are estimated to be 0.087 and 0.059, which are lower than the features of the Decision Tree model. Lastly, GBT model as a tree-base ensemble method is applied. Using 5-fold cross validation and grid search on hyperparameters, the results show that GBT improves the of model's performance better than RF with Adjusted R^2 of 0.896 for the test set while the RMSE and MAE are estimated to be 0.086 and 0.058.

Finally, we have compared the performance of these applied models, visually. Fig. 10 further shows the scatter plots between actual value and

Table 5
Results of ML algorithms.

Model		R^2	MAE	RMSE
Decision Tree	−ST_lag	0.579	0.106	0.135
	+ST_lag	0.797	0.064	0.094
Gradient-Boosted Tree	−ST_lag	0.674	0.084	0.101
	+ST_lag	**0.896**	**0.058**	**0.086**
Random Forest	−ST_lag	0.662	0.086	0.109
	+ST_lag	0.875	0.059	0.087

predicted value for each of models. As shown in Fig. 7, the closer predicted values are to the line, the better model fits the independent variable. Therefore, the GBT model are less scattered around the red line than other models. In addition, the residual scatter plot, which calculates the forecast error, shows that when using the GBT algorithm, the difference between the actual housing value and the predicted housing value is significantly less than the other used models, as shown on Fig. 11.

The results of these algorithms strongly prove that the used ST-lag variable boosts the performance of ML-based housing price prediction when there is spatiotemporal non-stationary in the data. Given that GBT model has better performance than other models, the feature importance is extracted using it. Fig. 12 ranked the top 10 features of the GBT model. The results show that ST-lag, dwelling area, IEO, and building age are the most important features among all incorporated features across the models. The importance of neighborhood-related factors confirms that the value of a property's value is determined by the values of nearby properties (Copiello, 2020). The key roles of the characteristics of the apartment include age, area and quality have been proven in several studies so far (Cao et al., 2019; Jha et al., 2020).

5. Conclusion

As noted in the United Nation's Sustainable Development Goal 11.1, housing accessibility and affordability is a critical consideration in the quality of life of those residing in urban environments (Sachs, 2012). Housing affordability is acknowledged as a key concern for those living in Australia Cities (Maclennan et al., 2021). Understanding property prices and those key characteristics which determine property value is an important consideration for those planning our cities and setting the dials on housing policy. This research helped improving our level of knowledge on housing value variations in an Australian city. The different types of variables in different scales such as property attributes and neighborhood quality can affect housing prices. Like, other key interested parties in understanding the contributing value of key variables contributing to property value includes real estate agent, developers, buys and investors, in order to assist in rational decisions regarding property sale and purchase.

In this paper, four ML models are trained with 32-year housing price data from metropolitan Adelaide utilizing multiple linear regression (MLR), Decision Tree (DT), Random Forest (RF), and Gradient-Boosted Tree (GBT). We have shown that ML algorithms can perform accurate prediction of housing prices, as evaluated by R2, RMSE, and MAE metrics. The results showed that GBT and RF can generate comparably better performance in housing price prediction with lower prediction errors in comparison with MLR and DT results. The comparison between the suite of ML models employed in this research showed that the relationships between housing price and features are non-linear. Therefore, non-linear tree-based regression models such as Decision Tree have better performance than the linear regression model. Moreover, ensemble ML techniques such as Gradient-Boosting and Random Forest are powerful methods to produce better predictive performance. Also, a proposed spatiotemporal lag (ST-lag) variable was incorporated successfully to improve the predictive accuracy of the models. Based on the findings of this study, we conclude that the ST-lag feature, which addresses spatiotemporal non-stationarity, can significantly outperform and improve the accuracy of ML-based housing price prediction models, and that changing the spatial lag bandwidth could improve the technique's applicability in general.

Another addition of this work is the use of ML as a predictive

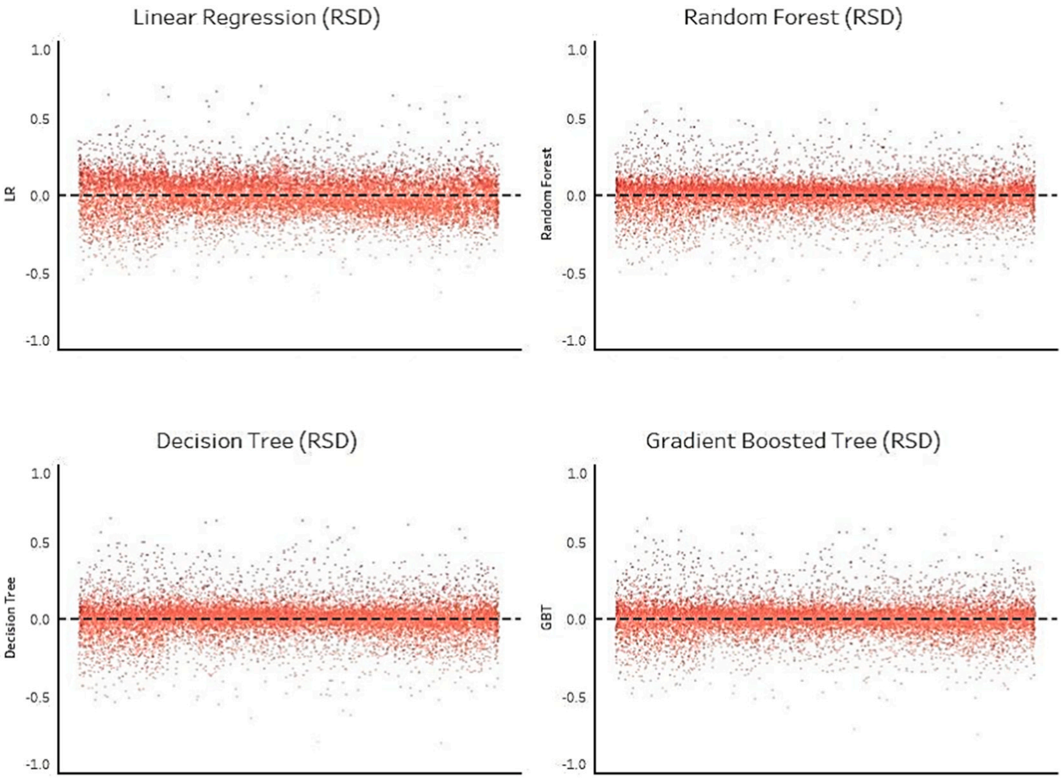


Fig. 11. Comparison of Models accuracy (scatter plot of model residuals).

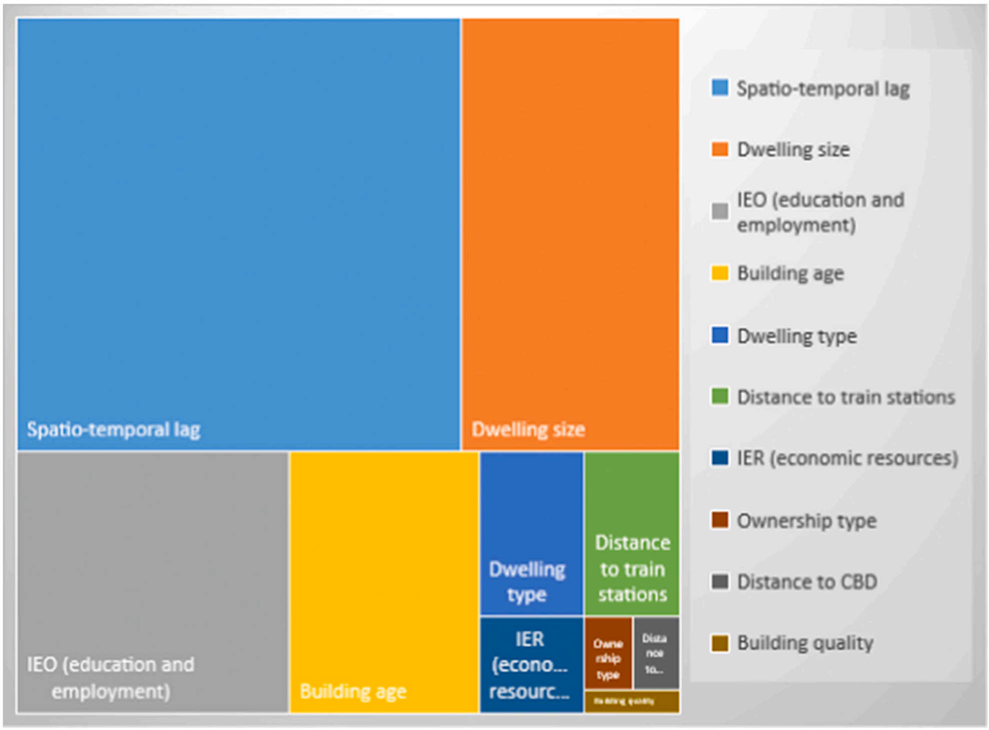


Fig. 12. Importance of top 10 features using GBT.

modelling strategy with high accuracy, which goes beyond the explanatory modelling approach commonly employed in the literature to provide causal explanations for housing prices (Ho et al., 2021; Shmueli, 2010). In reality, in the age of Big Data, one may learn in depth about the complicated behavior of the housing market for a huge number of input variables by employing cutting-edge ML technology as a subset of AI that leads to more logical decision making.

Our modelling technique and findings may help real estate investors, builders, property owners, house appraisers, and other stakeholders gain a more realistic view. This study, as one of the first of its kind, would have implications for policymakers by providing insights into the potential impacts of urban planning (such as infill regeneration, master planned communities, gentrification, and population displacement) and infrastructure provision policies on the housing market and subsequent local and regional economy. For example, the provision of downsized affordable housing in highly accessible public transit areas, as well as policy considerations to move metropolitan Adelaide away from the mono-centric structure seen in other capitals such as Sydney (Moghadam et al., 2017; Moghadam et al., 2018), would result in more balanced housing value across local governmental areas (LGAs) in the long run. Furthermore, providing greater incentives to LGAs to enhance built environment quality and expand access to basic services will result in more dispersed job and housing location options, resulting in a more balanced physical expansion of metropolitan Adelaide. Furthermore, the created model may be extended in the future to aid in the forecasting of future housing markets in scenarios including the dichotomy of infill versus suburban growth paradigms.

In summary, this paper seeks useful models for housing price projection while providing insights into the Australian housing market. This research can also be applied for transactional datasets of the housing market from different locations across Australia and overseas. This study, the first of its' type in South Australia, modifies and analyses spatio-temporal data from the housing market to better understand its complex and non-linear structure. The results of this study will can be used in a wide range of property economics applications, both theoretical and practical. The research's useful models will support advancements in real estate market planning, market investment management, tax and insurance evaluation, and property valuation methods. The model will aid in the development of a practical solution that consciously combines ML technology and housing data for mutually beneficial action. The research can add to the existing literature by demonstrating that ST-lag (or a similar sort of spatio-temporal indicator) can be an useful moderator of spatio-temporal effects in ML applications. This paper will serve as a springboard for future academic research into the dynamics of the South Australian property market, utilizing the benefits of cutting-edge technology to develop models for business and property valuation at various geographical levels.

The model will also aid in demonstrating proof of concept and will assist enhanced decision-making when valuing, pricing, and investing in the Australian real estate market; hence, this research will provide significant benefits to end users and relevant stakeholders. The findings may also be used to assist policymakers in making choices by allowing them to understand how planned operational, monetary, and fiscal actions such as quantitative easing, subsidies, and land use changes may affect real estate markets. Better policy development and management will also benefit regional and urban populations in South Australia.

However, there are various limitations in this article that should be considered in future research. To begin, various macroeconomic forces have affected the Australian property market, ranging from short range factors such as interest rate levels and investor behavior to longer range issues such as economic growth and population change (Berry & Dalton, 2004). This study does not take into account macroeconomic changes as external factors such as the inflation rate, stock market, GDP, and employment growth rate. For example, land transfer tax (stamp duty) affects not just house prices but also turnover (Leigh & Davidoff, 2011). A research of home prices in Australia from 1970 to 2003 showed that, in the long run, real disposable income and the consumer price index had a considerable and beneficial influence on real house prices. They are also heavily influenced, both positively and adversely, by the unemployment rate, real mortgage rates, equity values, and housing stock (Abelson et al., 2005). An equilibrium correction model of quarterly Australian house prices from 1972 to 2006 identifies the key long-run drivers as real non-property income per house, the proportion of working-age people, the unemployment rate, two government policy changes, real and nominal interest rates, and non-price credit conditions (Williams, 2009). In addition to housing-related rules and policies, planning limitations have a significant impact on supply elasticities; hence, historical patterns of land use and geography play a vital role in housing supply, influencing market conditions (Ball et al., 2010).

Secondly, temporal variations of some variables related to infrastructure, e.g. accessibility to urban services such as shopping centers, train lines, public spaces etc. are not considered, due to their low rate of spatial-temporal changes. Finally, advanced deep learning techniques should be examined as methods with potential for providing more in-depth time series predictions in order to attain greater performance and dependability in housing price prediction (Zhan et al., 2020). Deep learning may also incorporate subjective features of housing, such as landscape and place attachment. Given a picture as input, subjective qualities may be retrieved from property images. The model can thus mimic a human's ability to recognise and appreciate aesthetic value, as well as evaluate a property from several photos (Zhao et al., 2019). In order to create a stronger prediction model, these visual characteristics can be integrated with basic property attributes as explained in this work.

CRediT authorship contribution statement

Ali Soltani: Conceptualization, Management, Methodology, Data analysis, Writing, Validation.

Mohammad Heidary: Methodology, Scripting, Visualization, Coding.

Fatemeh Aghaei: Scripting, Software, Literature review, Original draft preparation.

Christopher Pettit: Supervision, Writing - Reviewing and Editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The author would like to thank UniSA Business, particularly Chris Leishman and Peter Rossini, for providing access to the real estate database. The authors would like to thank the editor of Cities and two anonymous reviewers for their helpful recommendations and constructive criticism in improving the quality of this work.

References

Abelson, P., Joyeux, R., Milunovich, G., & Chung, D. (2005). Explaining house prices in Australia: 1970–2003. *Economic Record, 81*, S96–S103.

Anselin, L. (2013). *Spatial econometrics: Methods and models. 4*. Springer Science & Business Media.

Ball, M., Meen, G., & Nygaard, C. (2010). Housing supply price elasticities revisited: Evidence from international, national, local and company data. *Journal of Housing Economics, 19*(4), 255–268.

Belke, A., & Keil, J. (2018). Fundamental determinants of real estate prices: A panel study of german regions. *International Advances in Economic Research, 24*(1), 25–45.

Berry, M., & Dalton, T. (2004). Housing prices and policy dilemmas: A peculiarly australian problem? *Urban Policy and Research, 22*(1), 69–91.

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

Cao, K., Diao, M., & Wu, B. (2019). A big data-based geographically weighted regression model for public housing prices: A case study in Singapore. *Annals of the American Association of Geographers, 109*(1), 173–186.

Chica-Olmo, J., Cano-Guervos, R., & Chica-Rivas, M. (2019). Estimation of housing Price variations using spatio-temporal data. *Sustainability, 11*(6), 1551.

Copiello, S. (2020). Spatial dependence of housing values in northeastern Italy. *Cities, 96*, Article 102444. https://doi.org/10.1016/j.cities.2019.102444

de la Luz Hernández-Flores, M., Otazo-Sánchez, E. M., Galeana-Pizaña, M., Roldán-Cruz, E. I., Razo-Zárate, R., González-Ramírez, C. A., & Gordillo-Martínez, A. J. (2017). Urban driving forces and megacity expansion threats. Study case in the Mexico City periphery. *Habitat International, 64*, 109–122. https://doi.org/10.1016/j.habitatint.2017.04.004

Diao, M., & Ferreira, J. (2010). Residential property values and the built environment: Empirical study in the Boston, Massachusetts. *Metropolitan Area. Transportation Research Record, 2174*(1), 138–147. https://doi.org/10.3141/2174-18

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Fotheringham, A. S., Crespo, R., & Yao, J. (2015). Geographical and temporal weighted regression (GTWR). *Geographical Analysis, 47*(4), 431–452.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism, 10*(2), 486.

Gupta, R., Marfatia, H. A., Pierdzioch, C., & Salisu, A. A. (2021). Machine learning predictions of housing market synchronization across US states: The role of uncertainty. *The Journal of Real Estate Finance and Economics*. https://doi.org/10.1007/s11146-020-09813-1

Harris, R., Dong, G., & Zhang, W. (2013). Using contextualized G eographically W eighted R egression to model the spatial heterogeneity of land prices in B eijingC hina. *Transactions in GIS, 17*(6), 901–919.

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

Jha, S. B., Babiceanu, R. F., Pandey, V., & Jha, R. K. (2020). *Housing market prediction problem using different machine learning algorithms: A case study*. arXiv preprint. arXiv: 2006.10092.

Kang, Z., Catal, C., & Tekinerdogan, B. (2020). Machine learning applications in production lines: A systematic literature review. *Computers & Industrial Engineering, 149*, Article 106773.

Karamujic, M. H. (2015). Housing: Why Is it important? In M. H. Karamujic (Ed.), *Housing affordability and housing investment opportunity in Australia* (pp. 8–45). London: Palgrave Macmillan UK.

Kiely, T. J., & Bastian, N. D. (2020). The spatially conscious machine learning model. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 13*(1), 31–49.

Leigh, A., & Davidoff, I. (2011). *How do stamp duties affect the housing market*. Real Estate Institute of Victoria. http://people.anu.edu.au/andrew.leigh/pdf/StampDuty.pdf.

Li, C., Zhao, J., & Xu, Y. (2017). Examining spatiotemporally varying effects of urban expansion and the underlying driving factors. *Sustainable Cities and Society, 28*, 307–320.

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. https://doi.org/10.1016/j.cities.2018.11.016

Lloyd Lawhon, L. (2009). The neighborhood unit: Physical design or physical determinism? *Journal of Planning History, 8*(2), 111–132.

Lock, O., Bain, M., & Pettit, C. (2021). Towards the collaborative development of machine learning techniques in planning support systems–a Sydney example. *Environment and Planning B: Urban Analytics and City Science, 48*(3), 484–502.

Ma, Y., & Gopal, S. (2018). Geographically weighted regression models in estimating median home prices in towns of Massachusetts based on an urban sustainability framework. *Sustainability, 10*(4), 1026.

Maclennan, D., Long, J., Pawson, H., Randolph, B., Aminpour, F., & Leishman, C. (2021). *Housing: taming the elephant in the economy*.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., & Owen, S. (2016). Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research, 17*(1), 1235–1241.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156.

Moallemi, M., & Melser, D. (2020). The impact of immigration on housing prices in Australia. *Papers in Regional Science, 99*(3), 773–786.

Moghadam, A. S., Soltani, A., & Parolin, B. (2017). Transforming and changing urban centres: The experience of Sydney from 1981 to 2006. *Letters in Spatial and Resource Sciences, 11*(1), 37–53. Springer.

Moghadam, A. S., Soltani, A., Parolin, B., & Alidadi, M. (2018). Analysing the space-time dynamics of urban structure change using employment density and distribution data. *Cities, 81*, 203–213.

Ngiam, K. Y., & Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology, 20*(5), e262–e273.

Nguyen, H. A., Soltani, A., & Allan, A. (2018). Adelaide`s east end tramline: Effects on modal shift and carbon reduction. *Travel Behaviour and Society, 11*, 21–30. https://doi.org/10.1016/j.tbs.2017.12.002

Pettit, C., Shi, Y., Han, H., Rittenbruch, M., Foth, M., Lieske, S., & Christensen, B. (2020). A new toolkit for land value analysis and scenario planning. *Environment and Planning B: Urban Analytics and City Science, 47*(8), 1490–1507.

Phan, T. D. (2018). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In *Paper presented at the 2018 international conference on machine learning and data engineering (iCMLDE)*.

Qian, X., & Ukkusuri, S. V. (2015). *Exploring spatial variation of urban taxi ridership using geographically weighted regression*. Retrieved from.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34–55.

Sachs, J. D. (2012). From millennium development goals to sustainable development goals. *The Lancet, 379*(9832), 2206–2211.

Salvati, L., Ciommi, M. T., Serra, P., & Chelli, F. M. (2019). Exploring the spatial structure of housing prices under economic expansion and stagnation: The role of sociodemographic factors in metropolitan Rome, Italy. *Land Use Policy, 81*, 143–152. https://doi.org/10.1016/j.landusepol.2018.10.030

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310.

Simlai, P. (2021). Predicting owner-occupied housing values using machine learning: An empirical investigation of California census tracts data. *Journal of Property Research, 1–32*. https://doi.org/10.1080/09599916.2021.1890187

Singh, A., Sharma, A., & Dubey, G. (2020). Big data analytics predicting real estate prices. *International Journal of System Assurance Engineering and Management*, 1–12.

Sipan, I., Mar Iman, A. H., & Razali, M. N. (2018). Spatial–temporal neighbourhood-level house price index. *International Journal of Housing Markets and Analysis, 11*(2), 386–411.

Smith, T. E., & Wu, P. (2009). A spatio-temporal model of housing prices based on individual sales transactions over time. *Journal of Geographical Systems, 11*(4), 33–39.

Soltani, A., Pettit, C. J., Heydari, M., & Aghaei, F. (2021a). Housing price variations using spatio-temporal data mining techniques. *Journal of Housing and the Built Environment*, 1–29.

Soltani, A., Allan, A., Khalaj, F., Pojani, D., & Mehdizadeh, M. (2021b). Ridesharing in Adelaide: Segmentation of users. *Journal of Transport Geography, 92*, 1–15. https://doi.org/10.1016/j.jtrangeo.2021.103030

Soltani, A., Allan, A., Javadpoor, M., & Lella, J. (2022a). Space syntax in analysing bicycle commuting routes in inner metropolitan Adelaide. *Sustainability, 14*(6), 3485. https://doi.org/10.3390/su14063485

Soltani, A., Allan, A., Pojani, D., Khalaj, F., & Mehdizadeh, M. (2022b). Users and non-users of bikesharing: how do they differ? *Transportation Planning and Technology, 45*. https://doi.org/10.1080/03081060.2021.2017215

Spark, A. (2021). *Machine learning library (MLlib) guide*. Apache Spark.

Streimikiene, D. (2015). Quality of life and housing. *International Journal of Information and Education Technology, 5*(2), 140.

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price prediction via improved machine learning techniques. *Procedia Computer Science, 174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111

Williams, D. M. (2009). *House prices and financial liberalisation in Australia. Economics series working papers 432*. University of Oxford, Department of Economics.

Wu, C., Ren, F., Hu, W., & Du, Q. (2019). Multiscale geographically and temporally weighted regression: Exploring the spatiotemporal determinants of housing prices. *International Journal of Geographical Information Science, 33*(3), 489–511.

Xu, R. (2015). *Machine learning for real-time demand forecasting*. Massachusetts Institute of Technology.

Yang, L., Liang, Y., Zhu, Q., & Chu, X. (2021). Machine learning for inference: Using gradient boosting decision tree to assess non-linear effects of bus rapid transit on house prices. *Annals of GIS, 1–12*. https://doi.org/10.1080/19475683.2021.1906746

Yao, J., & Stewart Fotheringham, A. (2016). Local spatiotemporal modeling of house prices: A mixed model approach. *The Professional Geographer, 68*(2), 189–201.

Yuan, F., Wei, Y. D., & Wu, J. (2020). Amenity effects of urban facilities on housing prices in China: Accessibility, scarcity, and urban spaces. *Cities, 96*, Article 102433. https://doi.org/10.1016/j.cities.2019.102433

Zhan, C., Wu, Z., Liu, Y., Xie, Z., & Chen, W. (2020). Housing prices prediction with deep learning: an application for the real estate market in Taiwan. In *, 2020. IEEE 18th international conference on industrial informatics (INDIN)* (pp. 719–724). https://doi.org/10.1109/INDIN45582.2020.9442244

Zhang, X., Huang, B., & Zhu, S. (2019). Spatiotemporal influence of urban environment on taxi ridership using geographically and temporally weighted regression. *ISPRS International Journal of Geo-Information, 8*(1), 23.

Zhao, Y., Chetty, G., & Tran, D. (2019). Deep learning with XGBoost for real estate appraisal. In *2019 IEEE symposium series on computational intelligence (SSCI)* (pp. 1396–1401). https://doi.org/10.1109/SSCI44817.2019.9002790

Zhou, Y. (2020). *Housing sale price prediction using machine learning algorithms*. Los Angeles: University of California.

Zhu, L., & Zhang, H. (2021). Analysis of the diffusion effect of urban housing prices in China based on the spatial-temporal model. *Cities, 109*, Article 103015. https://doi.org/10.1016/j.cities.2020.103015

Zolnik, E. (2021). Geographically weighted regression models of residential property transactions: Walkability and value uplift. *Journal of Transport Geography, 92*, Article 103029. https://doi.org/10.1016/j.jtrangeo.2021.103029

Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. (2020). House Price prediction using a machine learning model: A survey of literature. *International Journal of Modern Education & Computer Science, 12*(6).