BDMA_house_price_prediction

by Yaping Zhao

Submission date: 15-Nov-2023 05:03PM (UTC+0800)
Submission ID: 2099298632
File name: use_Price_Prediction__A_Multi_Source_Data_Fusion_Perspective.pdf (1.82M)
Word count: 6526
Character count: 37209

BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 0?/?? pp???–???
Volume 1, Number 1, January 2018

House Price Prediction: A Multi-Source Data Fusion Perspective

Yaping Zhao, Jichang Zhao, Edmund Y. Lam*

Abstract: House price prediction is of utmost importance in forecasting residential property prices, particularly as the demand for high-quality housing continues to rise. Accurate predictions have implications for real estate investors, financial institutions, urban planners, and policymakers. However, accurately predicting house prices is challenging due to the complex interplay of various influencing factors. Previous studies have primarily focused on basic property information, leaving room for further exploration of more intricate factors such as amenities, traffic, and social sentiments in the surrounding environment. In this paper, we propose a novel approach to house price prediction from a multi-source data fusion perspective. Our methodology involves analyzing house characteristics and incorporating factors from diverse aspects, including amenities, traffic, and emotions. We validate our approach using a dataset of 28,550 real-world transactions in Beijing, China, providing a comprehensive analysis of the drivers influencing house prices. By adopting a multi-source data fusion perspective and considering a wide range of influential factors, our approach offers valuable insights into house price prediction. The findings from this study have the potential to enhance the accuracy and effectiveness of house price prediction models, benefiting stakeholders in the real estate market.

Key words: price prediction; real estate; data mining; machine learning

1 Introduction

House price prediction plays a crucial role in the research area focused on forecasting residential property prices. As economic development progresses, the demand for higher quality housing has increased, underscoring the growing importance of accurate house price prediction. The implications of accurately predicting house prices extend to various stakeholders, including real estate investors, financial institutions, urban planners, and policymakers. Accurate predictions not only contribute to market surveillance but also empower sellers to determine optimal pricing strategies and assist potential buyers in making well-informed decisions.

However, accurately predicting house prices poses a considerable challenge due to the complex interplay of factors that influence them. The multifaceted nature of these factors makes it difficult to comprehensively measure and precisely predict house prices. Consequently, achieving comprehensive measurements and accurate predictions in house price prediction remains a formidable task.

Previous studies have predominantly focused on basic property information, such as the number of rooms and floor space area, to predict house prices [26]. However, more intricate factors, including the surrounding amenities, traffic conditions, and social sentiments, have received limited attention, leaving room for further exploration and research.

In this paper, we present house price prediction from a multi-source data fusion perspective. Our methodology involves studying the characteristics of the house, as well as considering factors such as

• Yaping Zhao and Edmund Y. Lam are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong SAR. E-mail: {zhaoyp, elam}@eee.hku.hk
• Jichang Zhao is with the School of Economics and Management, Beihang University, Beijing, 100191, China. E-mail: jichang@buaa.edu.cn
* To whom correspondence should be addressed.

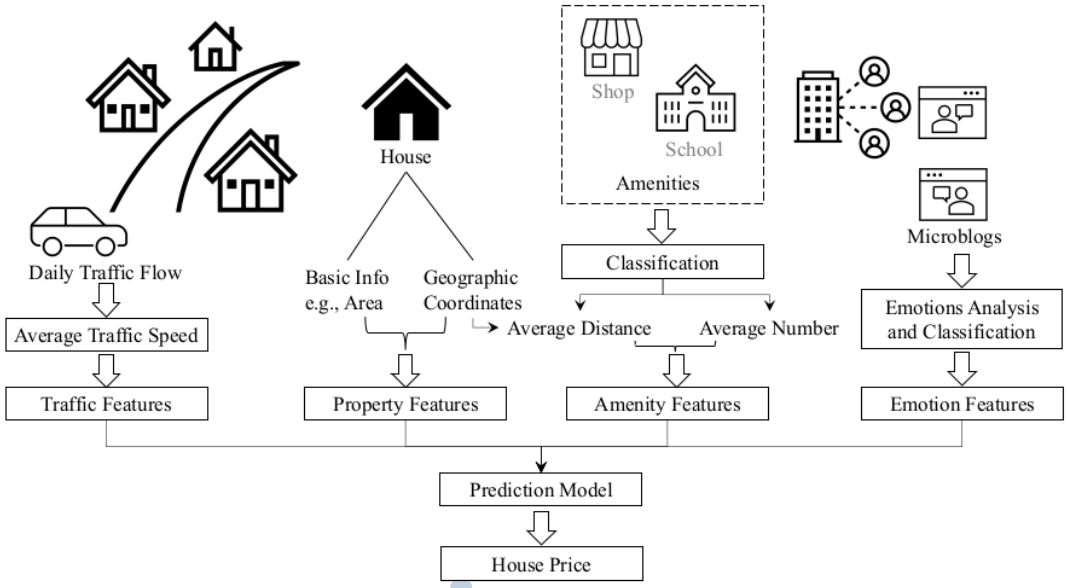

Fig. 1 The framework of the multi-source data fusion for house price prediction.

amenities, traffic, and emotions in the surrounding environment, as Figure 1 shows. To validate our approach, we conduct an analysis of 28,550 real-world transactions in Beijing, China, providing a comprehensive analysis of the drivers influencing house prices.

Our main contributions are as follows:

- We investigate house price prediction from a multi-source data fusion perspective and make a discovery that the characteristics from different aspects, such as amenities, traffic, and emotions, have an unexpected economic impact on house prices.
- We examine the correlation between various house features and their influence on house prices, ranking these features based on their importance.
- Through extensive experiments on a real-world dataset, we comprehensively compare and analyze the performance of different machine learning methods, including support vector machines, linear regression, XGBoost regression, and random forest regression, in the task of multi-source data fusion for house price prediction.

By adopting a multi-source data fusion perspective and considering a wide range of influential factors, our approach provides valuable insights into house price prediction. The findings from this study have the potential to enhance the accuracy and effectiveness of house price prediction models, ultimately benefiting various stakeholders in the real estate market.

The subsequent sections of this paper will delve into the related work (Sec. 2), methodology (Sec. 3), experiment (Sec. 4), and conclusions (Sec. 5), further elucidating our approach and presenting the implications of our findings.

2 Related Work

The house price prediction is often approached as the valuation of a heterogeneous good, characterized by a combination of utility-bearing features [8, 20]. Consequently, the price of a house can be seen as a quantitative representation of a set of these features. Numerous studies have been conducted over the past decades to explore the relationship between house prices and their associated features. For example, Krol [17] investigated the correlation between apartment prices and significant features using hedonic analysis in Poland. In Turkey, [29, 34] examined the positive and negative effects of different house features on house values. Kryvobokov and Wilhelmsson [18] determined the relative importance

Table 1 Comparisons with the existing house price prediction work. Our house data is more comprehensive in terms of house transaction records and house features than those used in the literature.

		References															Ours
		[3]	[17]	[34]	[29]	[27]	[2]	[4]	[23]	[8]	[18]	[25]	[33]	[19]	[1]	[12]	
#Data	>= 10,000	✘	✔	✘	✘	✔	✔	✔	✘	✔	✘	✘	✘	✘	✘	✔	✔
Property	Basic Info	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
	Geo-Info	✘	✘	✔	✔	✔	✔	✔	✘	✘	✔	✔	✘	✔	✔	✔	✔
Amenity	Transport	✘	✘	✔	✘	✘	✘	✘	✘	✘	✘	✔	✘	✔	✔	✔	✔
	Education	✘	✘	✔	✘	✔	✘	✘	✘	✘	✘	✘	✔	✔	✘	✔	✔
	Hospitals	✘	✘	✘	✘	✘	✘	✘	✘	✔	✔	✔	✔	✔	✘	✔	✔
	Shops	✘	✘	✔	✘	✘	✘	✘	✘	✔	✔	✔	✔	✔	✘	✔	✔
	Tourism	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✔
Traffic	Daily Flow	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✔
Emotion	Emotion	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✔

weights of location features that influence apartment market values in Donetsk, Ukraine. Ottensmann et al. [25] compared location measures, including distances and travel time to the central business district (CBD) and multiple employment centers, to understand the impact of residence location on house prices in Indianapolis, Indiana, USA. Ozalp and Akinci [33] identified housing and environmental features affecting residential real estate sale prices in Artvin, Turkey. These studies, along with many others, have investigated the relationship between house prices and various features, leading to the development of house price prediction approaches that estimate house prices based on inputted features.

Based on the underlying methodology, existing house price prediction methods predict house prices based on a range of constituent features and are typically applied directly to the entire dataset. Numerous studies have followed this approach. For instance, Gu et al. [14] leveraged support vector machines (SVM) [15] to predict house prices, showing promising results based on cases from China. Wang et al. [31] introduced a novel model based on SVM for predicting average house prices across different years, demonstrating the effective use of particle swarm optimization (PSO) for determining SVM parameters. Park and Bae [27] developed a general prediction model using machine learning techniques such as RIPPER, Naive Bayesian, and AdaBoost, comparing their classification accuracy performance.

However, these models often overlook the influence of house location and surroundings on prices, resulting in suboptimal prediction performance as the scale of the dataset increases. Recent studies have shifted towards local perspectives in house price prediction, serving as serious alternatives and extensions to traditional modeling approaches. Among these studies, Bourassa et al. [2] compared various methods for incorporating spatial dependence into house price prediction. Case et al. [4] demonstrated the importance of incorporating nearest neighbor transactions for accurate predictions. Gerek [13] designed two adaptive approaches, considering grid partition and sub-clustering. Montero et al. [23] considered model variants to capture spatial effects in house prices, proposing a mixed model that accounted for spatial autocorrelation, spatial heterogeneity, and nonlinearities. The results suggested that nonlinear models were more effective in house price prediction. Although some recent studies incorporate inputs such as infrastructure [11] and neighborhoods [6,7,10,16,30, 32], they rely on limited factors and ignore the complex intricacies associated with a multi-source data model.

Although the house price prediction problem has been extensively studied, our work diverges from most existing research in several aspects. Firstly, our house dataset, as illustrated in Table 1, encompasses a more comprehensive range of transaction records and house features compared to the datasets utilized in previous studies. This enables us to conduct a more in-depth exploration of the impact of various features on house prices and provides a greater understanding of the prediction problem. Secondly, we approach house price prediction from a unique perspective by incorporating

multi-source data fusion techniques. Through this approach, we have made a discovery, revealing that characteristics from diverse aspects, such as amenities, traffic, and emotions, unexpectedly exert an economic influence on house prices. This novel finding adds a new dimension to the understanding of the factors driving house price dynamics.

3 Methodology

To provide a multi-source data analysis and competently predict the house price per square meter, we first perform data collection and pre-processing in Section 3.1. After extracting 27 features from raw data, we compute feature correlations in Section 3.2. We then analyze the feature importance to house price in Section 3.3. Finally, we adopt two different methods for house price prediction in Section 3.4: SVM [15], linear regression [24], XGBoost [5], and random forest [21].

It is worth noting that our objective is to attempt to predict house prices by combining factors such as property characteristics, amenities, traffic, and social emotions. Instead of designing and implementing novel prediction methods, we focus on exploring the effectiveness of data fusion and its benefits. As demonstrated in previous work where multi-source data is used for house price prediction [7, 35], we adopt commonly used prediction models. The experimental results in Section 4 show that some of these methods have already achieved feasible performance and sufficiently evaluated the impact of different factors from multi-source data on house prices.

3.1 Data Collection and Pre-processing

To facilitate our house price prediction task from a multi-source data fusion perspective, we undertake comprehensive data collection and pre-processing procedures. This subsection presents the various steps involved in acquiring and preparing the data for analysis.

3.1.1 Property Data Collection

We initiate the data collection process by sourcing house transaction data from online platforms. For each house transaction, we extract pertinent features that reflect the fundamental characteristics of the properties. Specifically, we capture information such as the building year and the number of bedrooms. Additionally, we obtain the geographical coordinates by leveraging the address associated with each house. Overall, by collecting data from diverse online sources, we amass a dataset comprising $28,550$ house transactions [36], each accompanied by its corresponding basic property features.

3.1.2 Amenities Extraction

Leveraging the geographical coordinates obtained earlier, we exploit the capabilities of Baidu Maps to identify and categorize surrounding amenities within a kilometer radius of each house. These amenities are classified into five distinct types: transportation, tourist attractions, educational institutions, healthcare facilities, and restaurants. As part of the pre-processing stage, we compute two key features for each property: the total count and the average distance of each amenity type. These features provide valuable insights into the availability and proximity of amenities in the vicinity of the properties.

3.1.3 Traffic Data Acquisition

To incorporate the impact of transportation efficiency on house prices, we utilize Baidu Maps to obtain traffic-related information surrounding each house. For every property, we collect traffic speed data at five-minute intervals, ranging from 6 a.m. to 12 a.m. Subsequently, we compute the average traffic speed value, which serves as an additional feature capturing the transportation efficiency in the vicinity of the property. This feature provides a quantitative indicator of the accessibility and convenience associated with the location [38].

3.1.4 Emotional Sentiment Analysis

To incorporate the emotional aspect into our predictive model, we collect microblog posts from Beijing. We employ the social emotions analysis algorithm proposed in [9] to analyze the emotions expressed in each microblog post. Our analysis classifies each microblog post into one of five emotional categories: anger, detest (dislike), happiness, sadness, or fear. For every house, we calculate the percentage of different emotions associated with it, which serves as emotional sentiment features in our dataset.

In total, our data collection and pre-processing efforts result in the extraction of 27 features from the multi-source data. These features encompass various aspects, including property characteristics, surrounding amenities, traffic conditions, and emotional sentiments. Detailed information regarding the name and description of these extracted features is provided in Table 2. As our primary objective is to predict house prices, we designate the feature `Price` in Table 2 as

Table 2 The name and description of multi-source features we collected and extracted for the real estate data.

#	Category	Feature	Description
0	Property	Year	the building year.
1		Elvt	whether there is an elevator in the building.
2		RmNum	the number of bedrooms in the house.
3		HllNum	the number of living and dining rooms in the house.
4		KchNum	the number of kitchens in the house.
5		BthNum	the number of bathrooms in the house.
6		Lat	the latitude of the house.
7		Lng	the longitude of the house.
8	Amenity	TspNum	the number of surrounding transportation infrastructure.
9		TspDst	the average distance of surrounding transportation infrastructure.
10		AtrNum	the number of surrounding tourist attractions.
11		AtrDst	the average distance of surrounding tourist attractions.
12		EdcNum	the number of surrounding education and training institutions.
13		EdcDst	the average distance of education and training institutions.
14		HthNum	the number of surrounding healthcare infrastructure.
15		HthDst	the average distance of surrounding healthcare infrastructure.
16		RstNum	the number of surrounding restaurants.
17		RstDst	the average distance of surrounding restaurants.
18		RtlNum	the number of surrounding retail goods and services.
19		RtlDst	the average distance of surrounding retail goods and services.
20	Traffic	TrfV	the average value of daily traffic speeds.
21	Emotions	AgrPct	the percentage of anger in all emotions.
22		DstPct	the percentage of detestation in all emotions.
23		HppPct	the percentage of happiness in all emotions.
24		SadPct	the percentage of sadness in all emotions.
25		FeaPct	the percentage of fear in all emotions.
26		Price	the price per square meter of the house in Renminbi (RMB[a]).

[a]RMB is the legal currency of China.

the dependent variable y, while the remaining features are treated as independent variables x_i, where $i = 0, \ldots, 25$.

3.2 Feature Correlation

Understanding the relationships between different features is crucial in our analysis. We employ the Pearson's correlation coefficient, denoted as r_{uv}, to quantify the correlation between two features u and v [28]. The coefficient is calculated as follows:

$$r_{uv} = \frac{\sum_{j=1}^{n}(u_j - \overline{u})(v_j - \overline{v})}{\sqrt{\sum_{j=1}^{n}(u_j - \overline{u})^2}\sqrt{\sum_{j=1}^{n}(v_j - \overline{v})^2}}, \quad (1)$$

where n represents the sample size. The individual sample points of features u and v are denoted as u_j and v_j, respectively. The sample mean $\overline{x}$ is calculated as $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$, and a similar calculation is performed for $\overline{y}$.

The Pearson's correlation coefficient ranges between -1 and 1. A value closer to 1 indicates a strong positive correlation between the features, while a value closer to -1 suggests a strong negative correlation, implying that the features are more "opposite" in nature. A value close to 0 indicates a weak correlation between the features.

3.3 Feature Importance

The concept of feature importance is crucial in evaluating the usefulness and value of each feature in constructing boosted decision trees within the model. The more frequently an attribute is utilized in making critical decisions within decision trees, the higher its relative importance. By explicitly estimating the importance of different features, we can rank and compare them.

To estimate the importance of features in the context of predicting house prices, we utilize the random forest library in Python. This approach allows us to assess

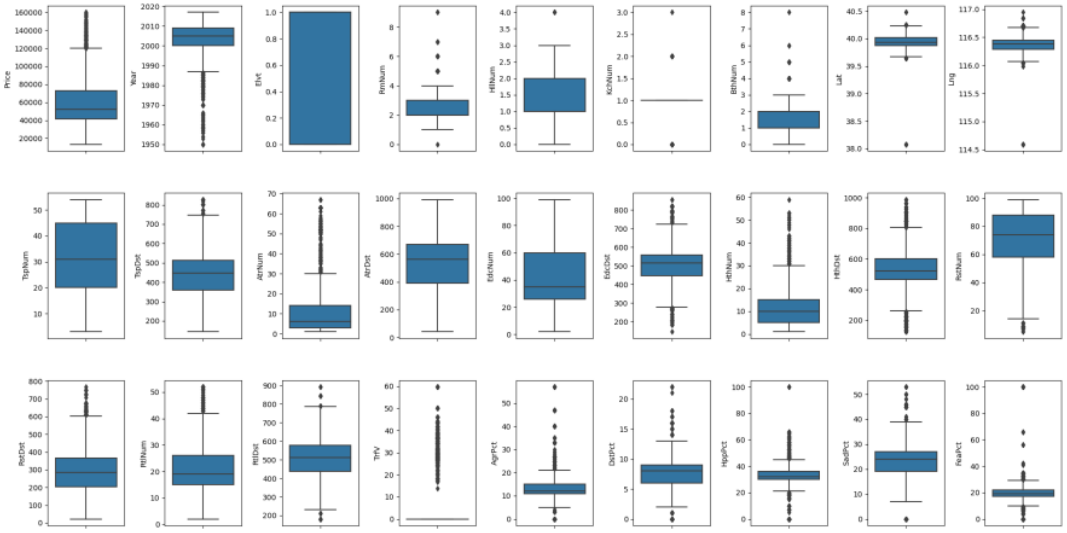

Fig. 2 The boxplots of the multi-source features with interesting trends/stats.

the relative importance of each feature and gain insights into their contributions to the predictive model.

3.4 Prediction Model

With the complex features derived from the multi-source data, we can construct prediction models for house prices. In our study, we employ four distinct methods: SVM [15], linear regression [24], XGBoost [5], and random forest [21]. Previous research [7, 22, 35] demonstrates that these commonly used regression models yield competent performance in the task of house price prediction.

By leveraging the diverse features obtained from multiple sources, we aim to develop robust models capable of accurately predicting house prices.

3.5 Evaluation

To measure the accuracy of our prediction models, we employ five evaluation metrics: R-squared (R^2), adjusted R^2, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). These metrics provide comprehensive insights into the performance of the models.

The formulas for these metrics are as follows:

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \overline{y})^2}, \tag{2}$$

$$\text{Adjusted } R^2 = 1 - \left[\frac{(1-R^2)\times(n-1)}{n-k-1}\right], \tag{3}$$

$$\text{MAE} = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{n}, \tag{4}$$

$$\text{MSE} = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n}, \tag{5}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n}}, \tag{6}$$

where n represents the sample size. The actual values of the dependent variable and their corresponding predicted values are denoted by y_j and $\hat{y}_j$, respectively. The average of all the actual dependent variable values is represented by $\overline{y}$. The term k denotes the number of independent variables in the model, excluding the constant term. These evaluation metrics enable us to assess the accuracy and performance of our prediction models.

4 Experiment

4.1 Feature Distribution

In Figure 2, the multi-source features are represented using boxplots, providing a visual depiction of their distributions. To further examine the distributions of the multi-source features, Figure 3 presents histograms.

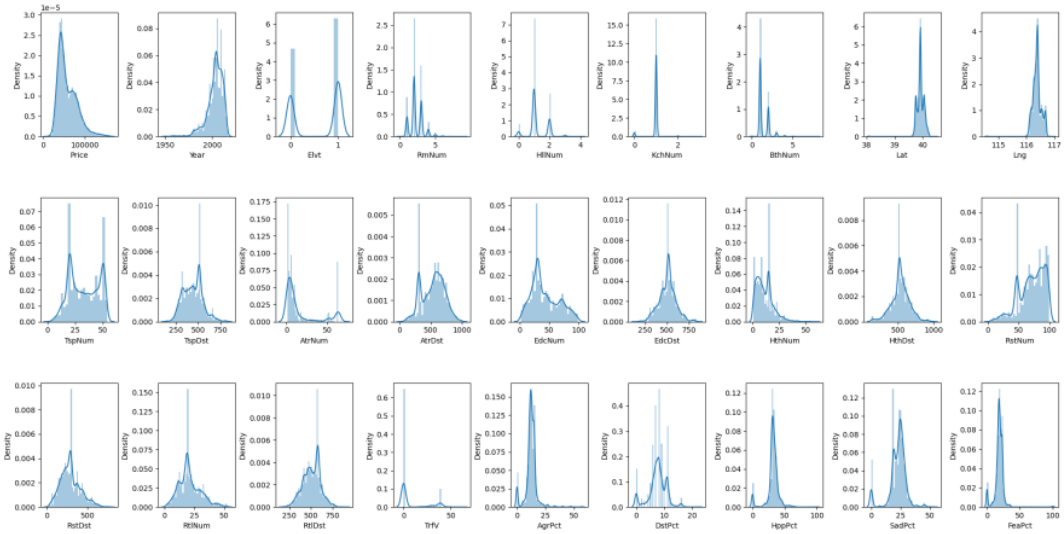

Fig. 3 The plots illustrate the distributions of the multi-source features. The histograms also show that features like `Price`, `Year` have highly skewed distributions. Also `EdcNum`, `EdcDst` look to have a normal distribution and other features seem to have norma or bimodel ditribution of data except `Elvt`, `RmNum`, `KchNum`, `BthNum` (which are discrete variables).

The histograms reveal certain characteristics of the features. Notably, features such as `Price` and `Year` exhibit highly skewed distributions. On the other hand, features like `EdcNum` and `EdcDst` appear to follow a normal distribution, while other features show either a normal or bimodal distribution of data, except for `Elvt`, `RmNum`, `KchNum`, and `BthNum`, which are discrete variables.

4.2 Feature Correlation

According to Figure 4, the absolute value of the correlation coefficient ($|r|$) between `Price` and other variables falls within the range of $[0, 0.5]$, indicating a weak linear correlation. This implies that relying solely on one or a few features is unlikely to accurately predict house prices. Instead, a collective gathering of multiple features from various sources is necessary and effective for accurate prediction.

Furthermore, in terms of the correlation with house prices, we also observe correlations between features within the same category. Specifically, the following interesting patterns are observed:

(1) Property features, such as `RmNum` (number of bedrooms) and `BthNum` (number of bathrooms), exhibit a correlation coefficient of 0.62, indicating a high correlation between the number of bedrooms and bathrooms in a house.

(2) Amenity features exhibit diverse relationships. For instance, the correlation coefficient between `EdcNum` (number of educational institutions) and `RstNum` (number of restaurants) is 0.77, between `RstNum` and `RtlNum` (number of retailing facilities) it is 0.74, and between `EdcNum` and `RtlNum` it is 0.68. These observations suggest that educational institutions often have restaurants and retailing facilities in close proximity. Moreover, restaurants and retailing services are frequently clustered together in the same area.

(3) The traffic feature `TrfV` shows a relatively strong relationship with `AtrNum` (number of tourist attractions), indicating that areas with a high number of tourist attractions tend to have high vehicular traffic.

Notably, the features `AtrNum`, `EdcNum`, `HthNum`, `RstNum`, `TrfV`, `DstPct` exhibit a higher correlation score with `Price`. To delve deeper into the relationship between each individual feature and the price, we present Figure 5. As depicted in Figure 5, while there is a faint hint of a linear fit in the overall trend of the data, it is not evident and can be considered negligible.

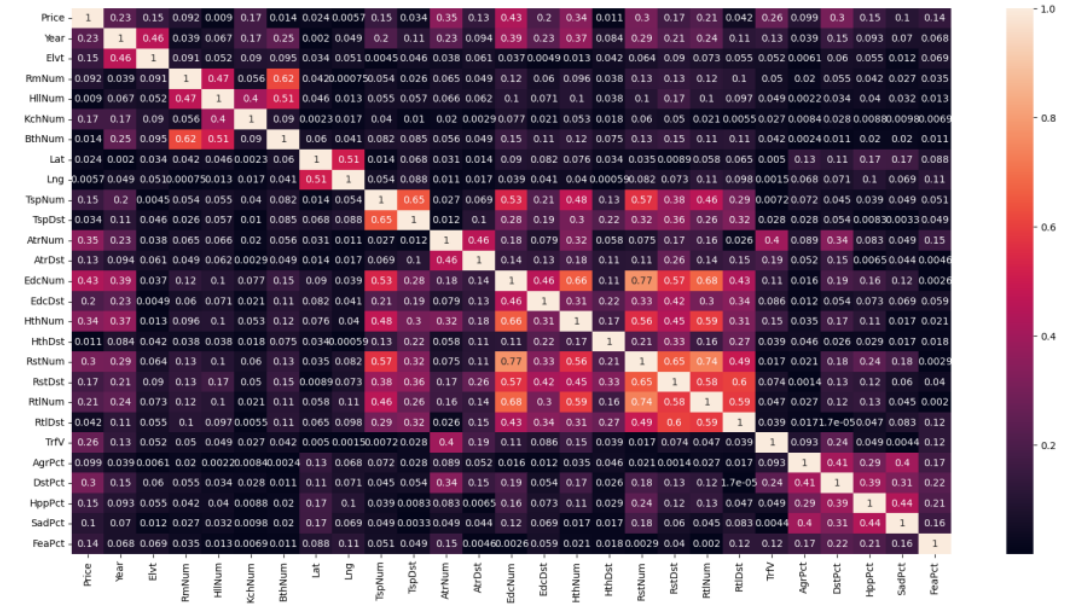

Fig. 4 The relationship between multi-source features, measured by the absolute of the r value, also called Pearson's correlation coefficient [28]. From this correlation matrix , we can see that if we only consider individual features, there is no significant correlation between any feature and `Price`. However, we can also observe that there may be strong correlations between some features (excluding `Price`). For instance, we see `RmNum` and `BthNum`, `EdcNum` and `RstNum` are highly correlated features.

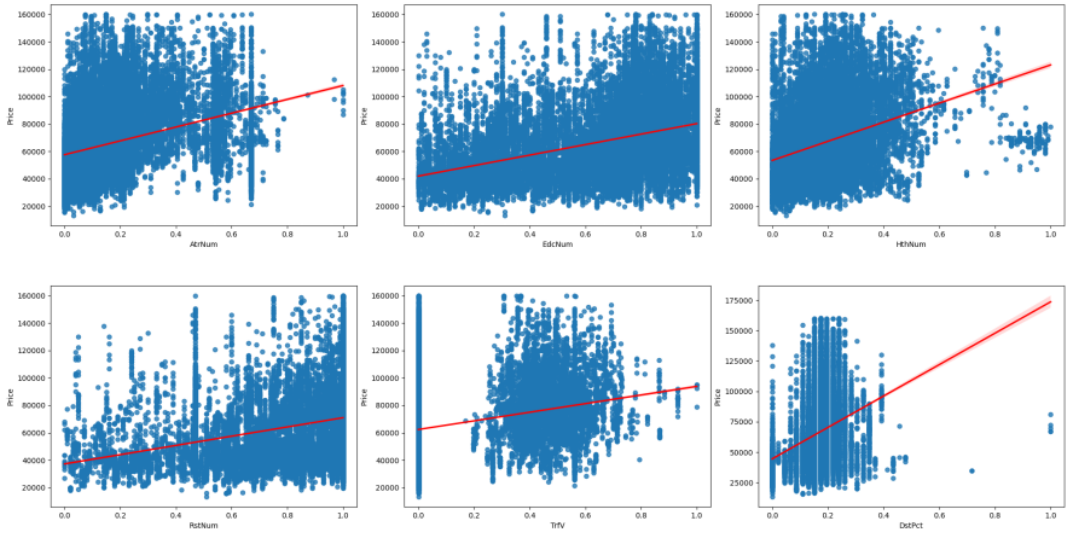

Fig. 5 The plots of the selected features against `Price`. where the features are with a higher correlation score with `Price` according to the correlation matrix in Figure 4.

4.3 Feature Importance

To assess the importance of different features in the random forest regression model, Figure 6 provides a ranking. The following observations can be made:

(1) The geographical location of the house, represented by its latitude (`Lat`) and longitude (`Lng`), emerges as the most influential factor affecting the price.

(2) Among all the features related to amenities, the average number of surrounding tourist attractions (`AtrNum`) and education institutions (`EdcNum`) prove to be the most significant. This implies that the availability of sightseeing opportunities and educational proximity are crucial considerations for potential house buyers.

(3) The year of construction (`Year`), which indicates the age of the house, ranks as the fifth most important factor influencing the price.

(4) The average distance to surrounding transportation infrastructure (`TspDst`) is the sixth most significant factor affecting the price.

(5) Interestingly, the average value of daily traffic speeds (`TrfV`) ranks 24th out of the 26 features. This suggests that while proximity to public transportation is valued by house consumers, the level of traffic congestion in the vicinity holds relatively less importance.

These findings shed light on the relative importance of various features in predicting house prices, providing valuable insights for real estate market analysis and decision-making.

4.4 Prediction Model

The dataset consists of $28,550$ data points, comprising a dependent variable y and 26 independent features x_i ($i = 0, \ldots, 25$). To train and evaluate our prediction models, we randomly split the dataset, allocating 70% of the data as the training set and the remaining 30% as the testing set.

We employ two different methods for house price prediction and analyze their outcomes as follows. After training, we obtain SVM, linear regression, XGBoost, and random forest models. To facilitate a comprehensive comparison among the linear, XGBoost, and random forest regression models, we present their results on the testing set visually in Figure 7. The following observations can be made:

Table 3 Evaluation and comparision of all the models.

Model	R	MAE	RMSE
Support Vector Machines	-0.5579	19833	25243
Linear Regression	36.5106	15235	20057
XGBoost	87.6973	5721	8829
Random Forest	89.3400	4932	8219

(1) Figures 7(a), 7(b), and 7(c) illustrate the disparities between the actual prices and the predicted values of the linear, XGBoost, and random forest regression models, respectively. While the predicted values of the linear regression model tend to be higher than the actual prices, the data points of the XGBoost and random forest models are evenly distributed around the $x = y$ line. This indicates that the predictions of the XGBoost and random forest models are more accurate.

(2) Figures 7(d), 7(e), and 7(f) display the residuals, representing the differences between the actual prices and the predicted values of the linear, XGBoost, and random forest regression models, respectively. The residuals of the XGBoost and random forest models are evenly distributed around zero, while the linear regression model exhibits numerous noticeable outliers, particularly around the predicted value of $80,000$.

(3) Figures 7(g), 7(h), and 7(i) depict the histograms of the errors for the linear, XGBoost, and random forest regression models, respectively. The errors in all models exhibit a normal distribution. However, the variance of the errors in the linear regression model is greater than that of the XGBoost and random forest models.

These comparisons provide valuable insights into the performance of different prediction models for house price estimation. Table 3 provides the evaluation and comparison of all the models.

4.5 Ablation Study

To assess the influence of different factors derived from multi-source data on house prices, we examine the performance of the SVM, linear regression, XGBoost, and random forest models on both the training and testing datasets.

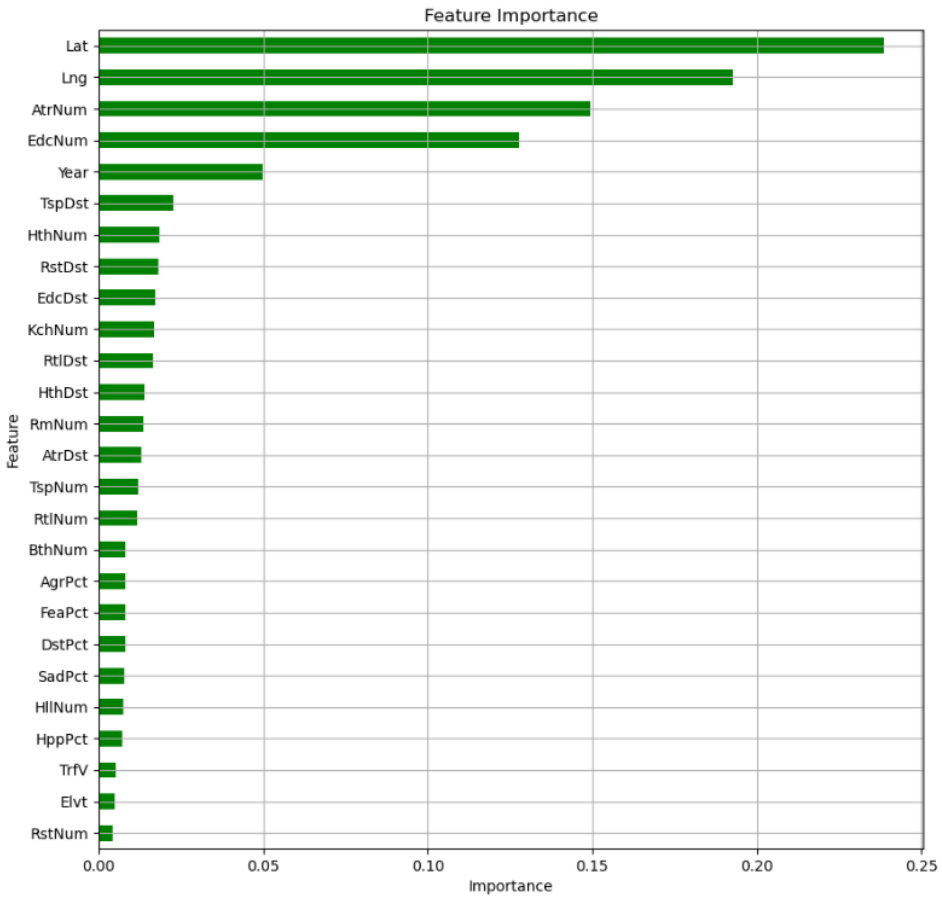

Fig. 6 The figure displays the ranking of features in the random forest regression model, assessing their importance in predicting house prices. Several key observations can be made from the ranking: 1) The geographical location of the house, represented by latitude (`Lat`) and longitude (`Lng`), emerges as the most influential factor. 2) Among amenity-related features, the average number of surrounding tourist attractions (`AtrNum`) and education institutions (`EdcNum`) are the most significant, highlighting the importance of sightseeing opportunities and educational proximity for potential buyers. 3) The year of construction (`Year`) ranks as the fifth most important factor, indicating its impact on prices. 4) The average distance to transportation infrastructure (`TspDst`) is the sixth most significant factor. 5) Surprisingly, the average value of daily traffic speeds (`TrfV`) ranks relatively low, suggesting that traffic congestion holds less importance compared to proximity to public transportation.

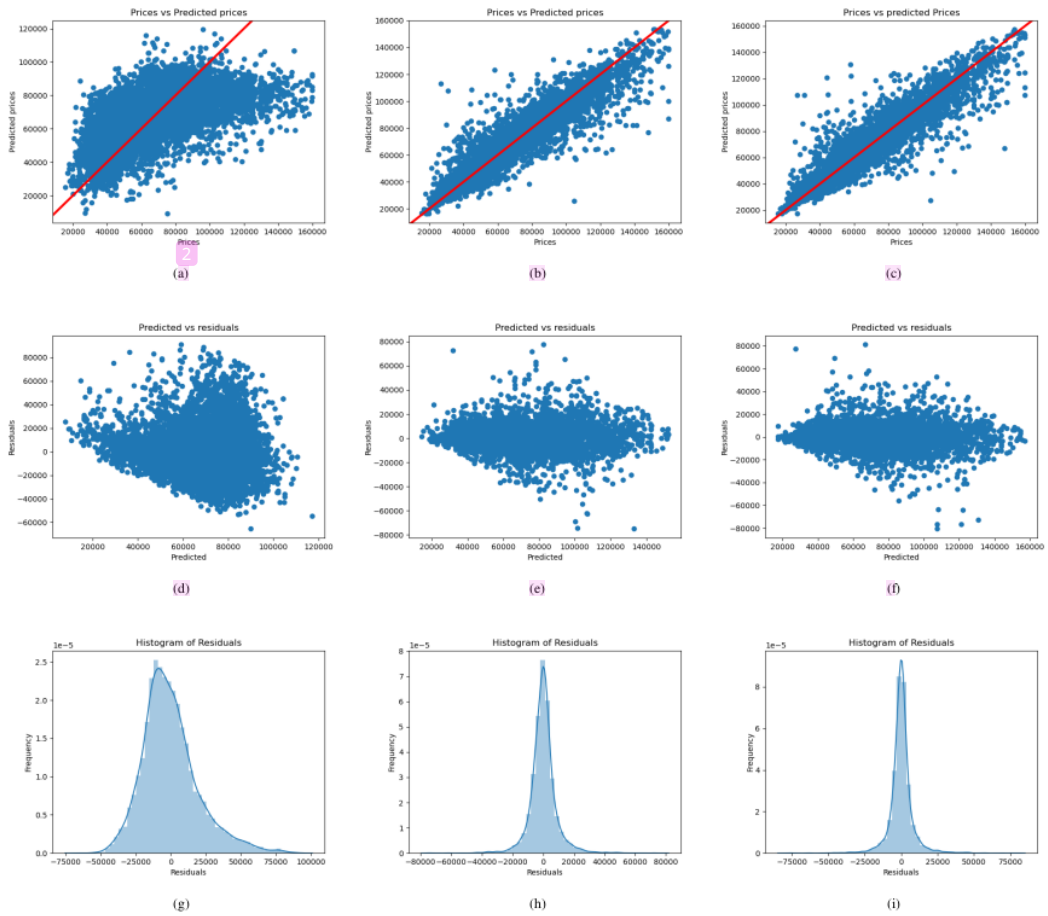

Fig. 7 Intuitive comparisons between the linear, XGBoost and random forest regression model for house price prediction. (a) The differences between actual prices and predicted values of the linear regression model. The red curve is the $x = y$ line. (b) The differences between actual prices and predicted values of the XGBoost regression model. (c) The differences between actual prices and predicted values of the random forest regression model. (d) The residuals of actual prices and predicted values of the linear regression model. (e) The residuals of actual prices and predicted values of the XGBoost regression model. (f) The residuals of actual prices and predicted values of the random forest regression model. (g) The histogram of errors of the linear regression model. (h) The histogram of errors of the XGBoost regression model. (i) The histogram of errors of the random forest regression model.

Table 4 Experiments with a variety of settings: 1) using only property features, denoted as *w/ only P*; 2) using property, traffic, emotions features but without amenity features, denoted as *w/o A*; 3) using property, amenity, emotions features but without traffic features, denoted as *w/o T*; 4) using property, amenity, traffic features but without emotions features, denoted as *w/o S*; 5) using all features including property, amenity, traffic, and emotions, denoted as *w/ PATE*.

Data	Method	$R^2 \uparrow$	Adjusted $R^2 \uparrow$	MAE $\downarrow$	MSE $\downarrow$	RMSE $\downarrow$
Training set	Support vector machines w/ only P	-0.0148	-0.0152	20251	672425564	25931
	Support vector machines w/o A	-0.0122	-0.0129	20219	670742089	25898
	Support vector machines w/o T	-0.0075	-0.0088	20177	667666607	25839
	Support vector machines w/o E	-0.0085	-0.0095	20189	668287594	25851
	Support vector machines w/ PATE	**-0.0082**	**-0.0095**	**20182**	**668044046**	**25847**
	Linear regression w/ only P	0.1674	0.1671	18284	551713399	23489
	Linear regression w/o A	0.2520	0.2515	16947	495668829	22264
	Linear regression w/o T	0.3730	0.3722	15499	415469247	20383
	Linear regression w/o E	0.3636	0.3629	15582	421702012	20535
	Linear regression w/ PATE	**0.3797**	**0.3789**	**15391**	**411032381**	**20274**
	XGBoost regression w/ only P	0.9095	0.9095	5206	59965069	7744
	XGBoost regression w/o A	0.9184	0.9183	4941	54069367	7353
	XGBoost regression w/o T	0.9331	0.9330	4416	44350499	6660
	XGBoost regression w/o E	0.9319	0.9318	4477	45145802	6719
	XGBoost regression w/ PATE	**0.9343**	**0.9342**	**4387**	**43549356**	**6599**
	Random forest w/ only P	0.9721	0.9721	2480	18466444	4297
	Random forest w/o A	0.9722	0.9722	2461	18372367	4286
	Random forest w/o T	0.9726	0.9725	2448	18143797	4259
	Random forest w/o E	0.9726	0.9726	2456	18135870	4265
	Random forest w/ PATE	**0.9726**	**0.9726**	**2448**	**18133433**	**4258**
Testing set	Support vector machines w/ only P	-0.0124	-0.0134	19899	641522761	25328
	Support vector machines w/o A	-0.0099	-0.0116	19871	639963349	25297
	Support vector machines w/o T	-0.0049	-0.0078	19827	636760159	25234
	Support vector machines w/o E	-0.0059	-0.0084	19840	637453688	25247
	Support vector machines w/ PATE	**-0.0056**	**-0.0086**	**19833**	**637188037**	**25243**
	Linear regression w/ only P	0.1626	0.1618	17905	530648157	23036
	Linear regression w/o A	0.2437	0.2424	16696	479250332	21892
	Linear regression w/o T	0.3591	0.3572	15324	406118449	20152
	Linear regression w/o E	0.3510	0.3494	15358	411213070	20278
	Linear regression w/ PATE	**0.3651**	**0.3632**	**15235**	**402302484**	**20057**
	XGBoost regression w/ only P	0.8560	0.8558	6244	91267181	9553
	XGBoost regression w/o A	0.8646	0.8644	6080	85802314	9263
	XGBoost regression w/o T	0.8751	0.8747	5773	79153827	8897
	XGBoost regression w/o E	0.8740	0.8737	5814	79830419	8935
	XGBoost regression w/ PATE	**0.8770**	**0.8766**	**5721**	**77956264**	**8829**
	Random forest w/ only P	0.8867	0.8866	5037	71763746	8471
	Random forest w/o A	0.8893	0.8891	4967	70132674	8374
	Random forest w/o T	0.8920	0.8917	4941	68382700	8269
	Random forest w/o E	0.8929	0.8926	4941	67855085	8237
	Random forest w/ PATE	**0.8934**	**0.8931**	**4932**	**67547377**	**8219**

In our experiments, we investigate various settings: 1) utilizing only property features (*w/ only P*); 2) incorporating property, traffic, and emotions features while excluding amenity features (*w/o A*); 3) including property, amenity, and emotions features while excluding traffic features (*w/o T*); 4) integrating property, amenity, and traffic features while excluding emotions (*w/o E*); 5) utilizing all features, encompassing property, amenity, traffic, and emotions (*w/ PATE*).

As presented in Table 4, compared to the scenario of using solely traditional property features (*w/ only P*), the addition of any supplementary feature (amenity, traffic, or emotions) demonstrates performance improvement. Furthermore, the following observations are noted: 1) The utilization of comprehensive features from multiple aspects, including property, amenity, traffic, and emotions, yields the highest performance. 2) The exclusion of amenity features results in the most significant performance decline, compared to the exclusion of traffic or emotions. This suggests that amenity features have a greater impact than traffic and emotions. 3) The removal of traffic features has a relatively modest impact on performance. This can be attributed to the one-dimensional nature of the traffic feature. Nevertheless, even the inclusion of this one-dimensional traffic feature contributes to performance enhancement.

These findings highlight the importance of incorporating various types of data, particularly amenity features, in accurately predicting house prices. The ablation study provides valuable insights into the relative contributions of different feature categories, facilitating a deeper understanding of the underlying factors influencing house prices.

5 Conclusions

In this paper, we presented a comprehensive analysis of house price prediction from a multi-source data fusion perspective. By incorporating property features, amenity data, traffic information, and social emotions, we aimed to uncover the underlying factors influencing house prices. Through extensive experiments, we demonstrated that the integration of various types of data improves the predictive performance of the models. The comprehensive consideration of property features, amenity data, traffic information, and social emotions yielded the highest predictive accuracy.

These findings can guide real estate buyers, sellers, and policymakers in making informed decisions. Moreover, it stirs and benefits social and economic analysis [37]. Future research can explore additional data sources and advanced modeling techniques to further enhance the accuracy of house price prediction models.

References

[1] Alastair S Adair, Jim N Berry, and W Stanley McGreal. Hedonic modelling, housing submarkets and residential valuation. *Journal of property Research*, 13(1):67–83, 1996.

[2] Steven Bourassa, Eva Cantoni, and Martin Hoesli. Predicting house prices with spatial dependence: a comparison of alternative methods. *Journal of Real Estate Research*, 32(2):139–160, 2010.

[3] Ayse Can. Specification and estimation of hedonic housing price models. *Regional science and urban economics*, 22(3):453–474, 1992.

[4] Bradford Case, John Clapp, Robin Dubin, and Mauricio Rodriguez. Modeling spatial and temporal house price patterns: A comparison of four models. *The Journal of Real Estate Finance and Economics*, 29:167–191, 2004.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[6] Joe Cortright. *Walking the walk: How walkability raises home values in US cities*. CEOs for Cities, 2009.

[7] Marco De Nadai and Bruno Lepri. The economic value of neighborhoods: Predicting real estate prices from the urban environment. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 323–330. IEEE, 2018.

[8] Gang-Zhi Fan, Seow Eng Ong, and Hian Chye Koh. Determinants of house price: A decision tree approach. *Urban Studies*, 43(12):2301–2315, 2006.

[9] Rui Fan, Jichang Zhao, Yan Chen, and Ke Xu. Anger is more influential than joy: Sentiment correlation in weibo. *PloS one*, 9(10):e110184, 2014.

[10] Yanjie Fu, Yong Ge, Yu Zheng, Zijun Yao, Yanchi Liu, Hui Xiong, and Jing Yuan. Sparse real estate ranking with online user reviews and offline moving behaviors. In *2014 IEEE International Conference on Data Mining*, pages 120–129. IEEE, 2014.

[11] Yanjie Fu, Hui Xiong, Yong Ge, Zijun Yao, Yu Zheng, and Zhi-Hua Zhou. Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking

and clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1047–1056, 2014.

[12] Guangliang Gao, Zhifeng Bao, Jie Cao, A Kai Qin, and Timos Sellis. Location-centered house price prediction: A multi-task learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–25, 2022.

[13] Ibrahim Halil Gerek. House selling price assessment using two different adaptive neuro-fuzzy techniques. *Automation in Construction*, 41:33–39, 2014.

[14] Jirong Gu, Mingcang Zhu, and Liuguangyan Jiang. Housing price forecasting based on genetic algorithm and support vector machine. *Expert Systems with Applications*, 38(4):3383–3386, 2011.

[15] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[16] Desislava Hristova, Luca M Aiello, and Daniele Quercia. The new urban success: How culture pays. *Frontiers in Physics*, page 27, 2018.

[17] Anna Król. Application of hedonic methods in modelling real estate prices in poland. In *Data Science, Learning by Latent Structures, and Knowledge Discovery*, pages 501–511. Springer, 2015.

[18] Marko Kryvobokov and Mats Wilhelmsson. Analysing location attributes with a hedonic model for apartment prices in donetsk, ukraine. *International journal of strategic property management*, 11(3):157–178, 2007.

[19] Michael Kuntz and Marco Helbich. Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. *International Journal of Geographical Information Science*, 28(9):1904–1921, 2014.

[20] Hakan Kuşan, Osman Aytekin, and İlker Özdemir. The use of fuzzy logic in predicting house selling price. *Expert systems with Applications*, 37(3):1808–1813, 2010.

[21] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[22] CH Raga Madhuri, G Anuradha, and M Vani Pujitha. House price prediction using regression techniques: a comparative study. In *2019 International Conference on Smart Structures and Systems (ICSSS)*, pages 1–5. IEEE, 2019.

[23] José-María Montero, Román Mínguez, and Gema Fernández-Avilés. Housing price prediction: parametric versus semi-parametric spatial hedonic models. *Journal of Geographical Systems*, 20:27–55, 2018.

[24] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[25] John R Ottensmann, Seth Payton, and Joyce Man. Urban location and housing prices within a hedonic model. *Journal of Regional Analysis and Policy*, 38(1), 2008.

[26] Elli Pagourtzi, Vassilis Assimakopoulos, Thomas Hatzichristos, and Nick French. Real estate appraisal: a review of valuation methods. *Journal of Property Investment & Finance*, 2003.

[27] Byeonghwa Park and Jae Kwon Bae. Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert systems with applications*, 42(6):2928–2934, 2015.

[28] Karl Pearson. Correlation coefficient. In *Royal Society Proceedings*, volume 58, page 214, 1895.

[29] Hasan Selim. Determinants of house prices in turkey: Hedonic regression versus artificial neural network. *Expert systems with Applications*, 36(2):2843–2852, 2009.

[30] Wenshan Wang, Su Yang, Zhiyuan He, Minjie Wang, Jiulong Zhang, and Weishan Zhang. Urban perception of commercial activeness from satellite images and streetscapes. In *Companion Proceedings of the The Web Conference 2018*, pages 647–654, 2018.

[31] Xibin Wang, Junhao Wen, Yihao Zhang, and Yubiao Wang. Real estate price forecasting based on svm optimized by pso. *Optik*, 125(3):1439–1443, 2014.

[32] E Washington and E Dourado. The premium for walkable development under land use regulations. *SSRN Electron. J*, 2018.

[33] Ayse Yavuz Ozalp and Halil Akinci. The use of hedonic pricing method to determine the parameters affecting residential real estate prices. *Arabian Journal of Geosciences*, 10:1–13, 2017.

[34] Rüştü Yayar and Derya Demir. Hedonic estimation of housing market prices in turkey. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, (43):67–82, 2014.

[35] Yaping Zhao, Ramgopal Ravi, Shuhui Shi, Zhongrui Wang, Edmund Y Lam, and Jichang Zhao. Pate: Property, amenities, traffic and emotions coming together for real estate price prediction. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE, 2022.

[36] Yaping Zhao, Shuhui Shi, Ramgopal Ravi, Zhongrui Wang, Edmund Y Lam, and Jichang Zhao. H4m: Heterogeneous, multi-source, multi-modal, multi-view and multi-distributional dataset for socioeconomic analytics in case of beijing. In *IEEE International Conference on Data Science and Advanced Analytics*. IEEE, 2022.

[37] Yaping Zhao, Zhongrui Wang, and Edmund Y Lam. Improving source localization by perturbing graph diffusion. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9. IEEE, 2022.

[38] Bodong Zhou, Jiahui Liu, Songyi Cui, and Yaping Zhao. Large-scale traffic congestion prediction based on multimodal fusion and representation mapping. In *IEEE International Conference on Data Science and Advanced Analytics*. IEEE, 2022.

BDMA_house_price_prediction

ORIGINALITY REPORT

47%	43%	43%	18%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.eee.hku.hk Internet Source	14%
2	deepai.org Internet Source	7%
3	arxiv.org Internet Source	4%
4	www.arxiv-vanity.com Internet Source	3%
5	www.researchgate.net Internet Source	2%
6	Guangliang Gao, Zhifeng Bao, Jie Cao, A. K. Qin, Timos Sellis. "Location-Centered House Price Prediction: A Multi-Task Learning Approach", ACM Transactions on Intelligent Systems and Technology, 2022 Publication	2%
7	vuir.vu.edu.au Internet Source	1%

8	Yaping Zhao, Ramgopal Ravi, Shuhui Shi, Zhongrui Wang, Edmund Y. Lam, Jichang Zhao. "PATE: Property, Amenities, Traffic and Emotions Coming Together for Real Estate Price Prediction", 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), 2022 Publication	1 %
9	Ramos, Bernardo. "Deep Learning for House Prices", Stanford University, 2022 Publication	1 %
10	export.arxiv.org Internet Source	1 %
11	onlinelibrary.wiley.com Internet Source	1 %
12	"Databases Theory and Applications", Springer Science and Business Media LLC, 2021 Publication	1 %
13	ah.lib.nccu.edu.tw Internet Source	1 %
14	link.springer.com Internet Source	1 %
15	www.irjmets.com Internet Source	1 %
	www.ncbi.nlm.nih.gov	

16	Internet Source	1 %
17	Submitted to Prince of Songkla University Student Paper	<1 %
18	dspace.cvut.cz Internet Source	<1 %
19	pure.royalholloway.ac.uk Internet Source	<1 %
20	Submitted to University of Melbourne Student Paper	<1 %
21	Antonio Lorenzo-Espejo, Alejandro Escudero-Santana, María-Luisa Muñoz-Díaz, Alicia Robles-Velasco. "Machine Learning-Based Analysis of a Wind Turbine Manufacturing Operation: A Case Study", Sustainability, 2022 Publication	<1 %
22	Submitted to University of Sydney Student Paper	<1 %
23	Submitted to University of Hong Kong Student Paper	<1 %
24	files.osf.io Internet Source	<1 %
25	dspace.spbu.ru Internet Source	<1 %
	theses.lib.polyu.edu.hk	

26	Internet Source	<1%
27	jurnal.itscience.org Internet Source	<1%
28	eprints.port.ac.uk Internet Source	<1%
29	purehost.bath.ac.uk Internet Source	<1%
30	www.diva-portal.org Internet Source	<1%
31	Nicolás Bettancourt, Cristian Pérez, Valeria Candia, Pamela Guevara et al. "Virtual tissue microstructure reconstruction across species using generative deep learning", Cold Spring Harbor Laboratory, 2023 Publication	<1%
32	dora.dmu.ac.uk Internet Source	<1%
33	dalspace.library.dal.ca Internet Source	<1%
34	Submitted to City University of Hong Kong Student Paper	<1%
35	Submitted to University of Edinburgh Student Paper	<1%

36	"Intelligent Computing and Optimization", Springer Science and Business Media LLC, 2023 Publication	<1 %
37	www.cepar.edu.au Internet Source	<1 %
38	aclanthology.org Internet Source	<1 %
39	repository.nwu.ac.za Internet Source	<1 %
40	tesi.luiss.it Internet Source	<1 %
41	www.mdpi.com Internet Source	<1 %
42	"Applied Data Science in Tourism", Springer Science and Business Media LLC, 2022 Publication	<1 %
43	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1 %
44	Maryam Heidari, Samira Zad, Setareh Rafatirad. "Ensemble of Supervised and Unsupervised Learning Models to Predict a Profitable Business Decision", 2021 IEEE International IOT, Electronics and	<1 %

Mechatronics Conference (IEMTRONICS), 2021
Publication

45	Seong-Hoon Cho. "Spatial variation of output-input elasticities: Evidence from Chinese county-level agricultural production data", Papers in Regional Science, 3/2007 Publication	<1 %
46	Lecture Notes in Computer Science, 2015. Publication	<1 %
47	Yaping Zhao, Shuhui Shi, Ramgopal Ravi, Zhongrui Wang, Edmund Y. Lam, Jichang Zhao. "H4M: Heterogeneous, Multi-source, Multi-modal, Multi-view and Multi-distributional Dataset for Socioeconomic Analytics in the Case of Beijing", 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), 2022 Publication	<1 %
48	Oluwaseun Daniel Akinyemi, Mohamed Elsaadany, Numair Ahmed Siddiqui, Sami Elkurdy et al. "Machine learning application for prediction of sonic wave transit time - A case of Niger Delta basin", Results in Engineering, 2023 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		