Journal of Economics and Management Sciences; Vol. 8, No. 2; 2025
ISSN 2576-3008 E-ISSN 2576-3016
https://doi.org/10.30560/jems.v8n2p117

Gradient Boosting Decision Tree for House Price Prediction with Google Trends

Faye F.F. Jiang[1]

[1] Scholar, Hong Kong

Correspondence: Faye F.F. Jiang, Scholar, Hong Kong.

Received: March 15, 2025; Accepted: April 1, 2025; Published: April 5, 2025

Abstract

Predicting house price accurately can reflect the popularity of the housing market and help making decisions for investors and policymakers. Statistics of macro factors are commonly used for house price forecasting; however, macro factors obtained from government reports have defect of time lag and may impair the prediction performance. Google Trends data can serve as a leading sentiment indicator of people's attitudes and expectations toward the housing market and help improve house price prediction. Therefore, this study proposes a new methodology framework for house price prediction with Google Trends data. Recursive Feature Elimination (RFE), a feature selection method, is utilized to remove noisy data and improve feature quality. Gradient Boosting Decision Tree (GBDT) is adopted to establish models for house price forecasting. Real estate-related Google Trends data, along with the fundamental house price index (HPI) data are collected to predict the growth rate of HPI in the United States. Results show that RFE can effectively remove irrelevant features and improve the model performance. GBDT has higher and more stable prediction accuracy than other prediction models, especially when the predicted time span is long. Compared with models including fundamental HPI data only, models containing Google Trends data can exhibit higher and more stable prediction accuracy for long time span forecasting. Three categories of Google Trends indices, including "house rent", "housing market & real estate market", and "mortgage & real estate agency" are found to be the most important indicators of the variation of HPI growth rate.

Keywords: house price prediction, growth rate of house price index, Google Trends, gradient boosting decision tree, recursive feature elimination

1. Introduction

House price plays an important role in modern society [1]. House price can influence not only the real estate market, but also the consumption level of local civilian, land finance income of local government, and even the economic development level of a region and a country [2–4]. A stable and healthy housing market will improve the economic development, while a stagnant or overheating housing market could threaten social stability [5,6]. Since the Great Recession of 2008, which was triggered by the bursting of the U.S. housing bubble [7], more scholars have conducted studies on house price. Their research directions can be briefly divided into three aspects, including the influential factors of house price [8], the effectiveness of control policies on house price [9], and house price prediction [10]. Especially, house price prediction is always a hot research topic in both academia and house market. Predicting house price accurately can present the developing housing trend and help making decisions for investors and policymakers [11,12]. House buyers can make reasonable investment decisions and governments can reallocate housing resources according to the predicted house price.

Many scholars have conducted research on house price prediction. For example, Varma et al. [13] used weighted mean of various regression techniques to predict house prices. The weighted mean model obtained minimum error than individual algorithm on house price prediction. Wei and Cao [14] used dynamic model averaging (DMA) approach to predict house price change in China and they concluded that DMA obtained better prediction performance than traditional models (e.g., Bayesian model). However, most existing studies adopted statistical methods for house price foresting (e.g., DMA approach [16], the multivariate probit model [17], and the auto-regression model [11]). The statistical methods are developed with predefined assumptions (e.g., linearity, data distribution). If the specific assumptions are not fulfilled in practical house price datasets, the prediction performance may be affected. Therefore, this study aims to use a non-linear machine learning method namely Gradient Boosting Decision Tree (GBDT) to improve the performance of house price prediction [1,6,11].

Another limitation of existing studies is that most studies forecasted the house price using statistics of macro factors such as economy, industry, demography and policy. For example, Wang et al. [8] collected eight influential factors such as land price, development costs, the profits of developers, and economic development trend to forest house price change. However, macro factors obtained from government reports have defect of time lag. For example, the data for a given month/year are usually released in the next month/year, and they may be revised later. Therefore, the available data cannot reflect the current level of macro activity and this may impair the performance of house price prediction. Besides, these macro factors are objective statistics of social activities, while house price is mainly determined by supply and demand (subjective behavior). Therefore, macro factors cannot fully explain the complex variation of house price [15].

The lucky thing is that Google Trends can address the data limitations. Google Trends provides daily reports on the volume of queries related to various activities. Google Trends data provide an insight into the top activities people are interested in, and serve as a leading sentiment indicator of people's attitudes and expectations toward the specific activity [16]. A few researchers have adopted Google Trends index to study other economic activities. For example, Choi and Varian [16] concluded that models with google trends variables were superior than models without these predictors, such as a 21.5% increase in sale prediction of motor vehicles and a 13.6% increase in prediction of unemployment benefits. The existing research has shown that internet search volume can help improve performance of economic activity prediction. However, few studies have used Google Trends for house price prediction. Wei and Cao [16] used Google search index as an additional predictor, along with the traditional macroeconomic indicators to predict house price change in 30 major Chinese cities. However, they used only one search indicator of "house price", which may limit the contribution of Google Trends data on house price prediction.

The limitations of existing studies on house price prediction are summarized as follows: Firstly, previous studies only used a few of Google Trends features, which cannot fully utilize the power of Google Trends data on house price foresting. Secondly, most existing studies adopted statistical methods for house price foresting, which are based on predefined assumptions (e.g., linearity, data distribution). If the specific assumptions are not fulfilled in practical house price datasets, the prediction performance may be affected. Thirdly, previous studies seldom conducted feature selection because the collected macro features are very limited. This study aims to collect a large number of Google search indices related to house price; however, one potential risk is that noise and unimportant features will also increase. Therefore, a proper feature selection method should be implemented to eliminate the irrelevant features and improve prediction performance.

Therefore, this paper proposes a new methodology framework for house price prediction with Google Trends data. Recursive Feature Elimination (RFE), a feature selection method, is utilized to remove noisy data and improve the quality of the Google Trends features. A non-linear machine learning method namely Gradient Boosting Decision Tree (GBDT) is adopted as the major algorithm for house price forecasting. The proposed framework is applied for house price prediction in the United States. Month-over-month (MoM) growth rate of house price index (HPI) is selected as prediction target. Historical HPI data and 29 real-estate-related Google search indices, including information of real estate industry, macro-economy, mortgage, employment and others are collected as model inputs. Prediction performance of the proposed framework is well evaluated and the most important Google search indices related to HPI growth rate are uncovered and analyzed.

The structure of the paper is organized as follows: Section 2 introduces the methodology framework; Section 3 presents the case study; Experimental results are analyzed in Section 4 and Section 5 concludes the work.

2. Methodology Framework

Figure 1 presents the methodology framework proposed in this paper. It consists of three parts, data collection and preprocessing, model application and discussion. Firstly, house price index (HPI) data and real-estate-related Google Trends indices are collected. These data are well preprocessed to construct time-series samples. The Gradient Boosting Decision Tree-based Recursive Feature Elimination (GBDT-RFE) model is then applied to select the most important Google Trends features. The selected features are then input into the prediction models. Models with different algorithms and models with different time spans are evaluated and compared. Finally, feature importance are calculated by GBDT method and the most important Google Trends features related to HPI growth rate are analyzed.

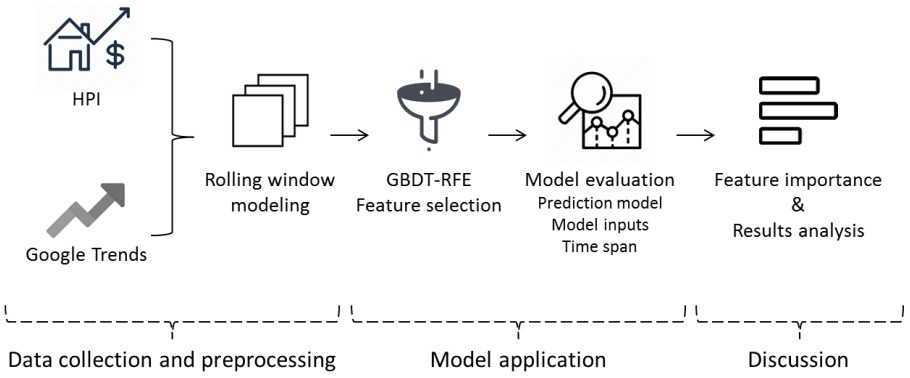


Figure 1. Methodology framework

2.1 GBDT Algorithm

2.1.1 Classification and Regression Tree

As shown in Figure 1, GBDT is the major prediction model adopted in this paper. GBDT is an ensemble learning method developed based on Classification and Regression Tree (CART) and the Gradient Boosting process [17–20].

CART is one kind of decision trees which selects the best split attribute and split node of each tree based on the variance of the sum [21]. For an arbitrary split attribute *j* and its corresponding split node *s*, the given dataset $R = \{x_i, y_i\}, 1 \le i \le N$ can be split into dataset R_1 and R_2. When the mean square error of R_1 and R_2 and their sum reach the smallest values, *j* can be considered as the best split attribute and *s* as the best split node. Mathematically, this process can be expressed as Equation (1).

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i - c_2)^2\right] \quad (1)$$

where c_1 represents the output mean value of R_1, c_2 represents the output mean value of R_2. The output of the node and the final output of the tree can be expressed as Equation (2) and Equation (3), respectively.

$$\hat{C}_m = \frac{1}{N_m}\sum_{x_i\in R_m(j,s)} y_i \ , x \in R_m \ , m = 1,2 \quad (2)$$

$$f(x) = \sum_{m=1}^{M} \hat{C}_m I(x \in R_m) \quad (3)$$

2.1.2 Gradient Boosting

Single CART usually has the problems of overfitting and weak generalization. To overcome the problems, GBDT combines multiple CARTs using Gradient Boosting [22]. Specifically, GBDT takes the minus gradient of the loss function of the last CART as the pseudo-residual and fits the current CART to the pseudo-residual. It keeps building trees to fit the residual until the aggregation of the loss function is minimized. Mathematically, for the *i*th sample on the *(m-1)*th tree, the pseudo-residual of its loss function can be expressed as Equation (4).

$$r_{im} = -\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \quad (4)$$

The output of the final leaf node of the *m*th tree then can be expressed as Equation (5).

$$c_{mj} = \underset{c_1}{argmin}\sum_{x_i\in R_{mj}} L(y_i, f_{m-1}(x_i) + c) \quad (5)$$

The learning model can be formulated as Equation (6).

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I(x \in R_{mj}) \quad (6)$$

The final CART given by GBDT is shown in Equation (7).

$$\hat{f}(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I(x \in R_{mj}) \quad (7)$$

2.1.3 Feature Importance

Except the great prediction ability of GBDT, the ability of GBDT in measuring feature importance is also utilized in this paper [12,19,20,23]. GBDT measures the feature importance based on the decrement of mean square error when each attribute is split [24]. For attribute *j*, its importance in one tree can be calculated by Equation (8).

$$\hat{J}_j^2(T) = \sum_{t=1}^{L-1} \hat{\imath}_t^2 1(v_t = j) \quad (8)$$

where L represents the number of leaf nodes and v_t represents the variables related with node *t*. Importance of *j* in all the M trees in GBDT can be expressed as Equation (9).

$$\hat{J}_j^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{J}_j^2(T_m) \quad (9)$$

2.2 RFE Algorithm

Compared with previous studies which utilized Google Trends data (e.g., house price, house rent) for house price prediction, real-estate-related Google search volume data collected in this paper are in a larger number and more comprehensive [20,23]. However, among the collected data, some of them might be redundant. They will not only slow down the calculation process but also lead to overfitting and lower the modeling performance. To remove these noisy features and ensure the modeling performance, a feature selection method named RFE is implemented in this paper.

RFE is a recursive process which aims at removing the weakest features until the pre-set number of features is reached and utilizing the remained features to fit a base model [25]. Whether the feature is weak or not is determined based on the coefficients or feature importance calculated by the base model [26,27]. By recursively eliminating a feature or a group of features per iteration, RFE eliminates dependencies and collinearity that may exist in the model. Pseudo code of the RFE algorithm is shown in Table 1. In this paper, GBDT is adopted as the rank model of RFE.

Table 1. Pseudo code of RFE

Algorithm 1: RFE algorithm
input： training dataset feature sets F; model step m process： for I in range(1:n+1, m): rank F by model fl = the last m important features Rank[int(n/m)-i+1] = fl pop fl from F Output: Rank

Overall, in this paper, a model based on GBDT and RFE is proposed for house price prediction and feature analysis. The model considers not only the historical house price data as the prediction indicator, but also a large number of real-estate-related Google Trends data for prediction. The data are well preprocessed and selected. To verify the effectiveness of the proposed methodology, a case study is conducted in Section 3.

3. Case Study

3.1 Data Collection

3.1.1 Growth rate of house price

The United States is selected as the study area in this paper due to the data availability. This study collects the monthly American HPI data between 2004 and 2017 provided by Freddie Mac [28] for experiments. The MoM growth rate of HPI are calculated in Equation (10). Calculation formulation is presented in Equation (10).

$$gr_t = \frac{hpi_t}{hpi_{t-1}} - 1 \tag{10}$$

where gr_t indicates the growth rate of HPI at time t; hpi_t and hpi_{t-1} donate the HPI at time t and $t-1$ separately.

168 samples of HPI growth rate during the study period is presented in Figure 2. It can be observed that the HPI growth rate follows a certain variation cycle. During one year, HPI growth rate first increases month by month, reaches its peak in the middle of the year, and then begins to fall. It can also be observed that the HPI growth rate kept decreasing from 2004 to 2008 and shows the first negative growth in 2006. The HPI growth rate reached the lowest point at the end of 2008 due to the bursting of the US housing bubble and the great recession. After 2009, the HPI growth rate began to rebound and became positive in 2013. The variation of the growth rate showed a stable trend during 2013 and 2017.

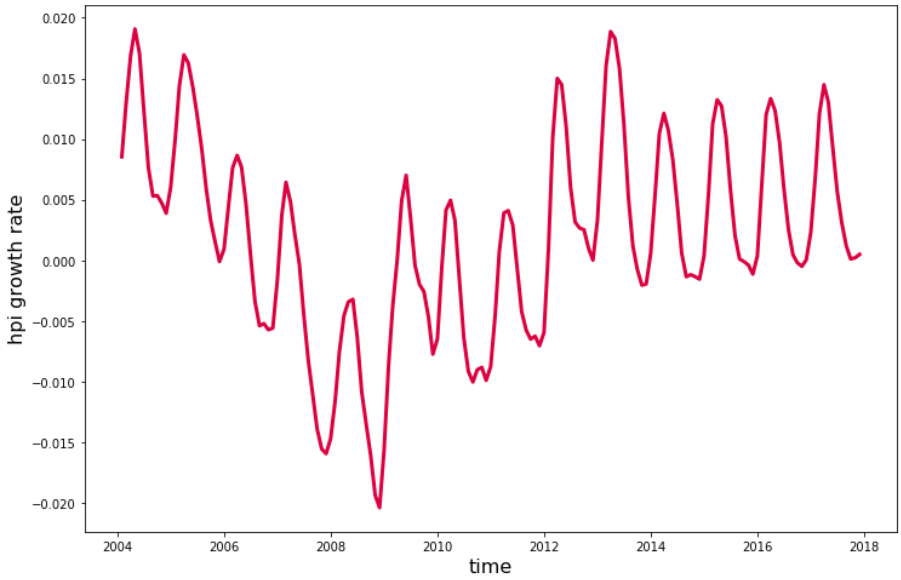


Figure 2. The MoM growth rate of HPI from 2004 to 2017

3.1.2 Google Trends Data

Google Trends is a website providing an index of the volume of Google queries by geographic location and category [29]. Considering the variation of the overall Google queries in different years, the query share, instead of the real search volume is firstly calculated. The query share is the ratio of the term's search volume to the sum of the total search volume in a given region at a given time. The values of the query share are then normalized to calculate the values of query index to indicate the popularity of the search term.

Google Trends data are collected as predictors for house price forecasting in this paper. To comprehensively explore the effect of Google Trends data on HPI growth rate prediction, five aspects of economic activities are considered, including real estate industry, macro-economy, mortgage, employment and others. As is shown in

Table 2, 29 key words (search terms) related to the five aspects are defined for HPI growth rate foresting. It is worth noting that Google Trends adopts an automated classification engine to include 27 sub-categories for each search term. This study selects six sub-categories, which are most related to housing prices, including "all categories", "real estate", "business & industrial", "finance", "law & government", and "jobs & education". Based on this fact, the feature dimension of Google Trends data at a given time point is 174 (29*6=174).

Table 2. Google Trends indices

Aspects	**key words**
Real estate industry	house price
	house rent
	housing market
	real estate market
	land price
	land value
	real estate agency
Macro-economy	economic
	financial crisis
	economic crisis
	economic growth
	cpi
	stock
	gdp
	income
	interest rate
Mortgage	mortgage
	bank
	loan
Employment	unemployment
	employment
	job
	hire
	recruitment
	salary
Others	population
	immigration
	federal fund
	invest

3.2 Data Preprocessing

The collected HPI growth rate data and Google Trends data are in time-series [20,23]. This study adopts the rolling window (RW) method to construct time-series samples for HPI growth rate prediction [30]. The form of time series samples is expressed as Equation (11)

$$\hat{y}_{t+h} = f(x_{t-1}, x_{t-2}, \dots, x_{t-n}, g_{t-1}^{1}, g_{t-2}^{1}, \dots, g_{t-n}^{1}, \dots, g_{t-1}^{k}, g_{t-2}^{k}, \dots, g_{t-n}^{k}) \quad (11)$$

where $\hat{y}_{t+h}$ indicates the predicted HPI growth rate at time $t+h$; h represents the time span for prediction; x_{t-i} represents the value of the HPI growth rate at time t-i; k represents the total number of Google Trends features; g_{t-i}^{l} represents the popularity of the lth Google Trends feature i months ago; and n represents the value of rolling window size or time lag.

Since the variation of the HPI growth rate follows an annual cycle and in order to consider enough information in model establishment, this paper sets the window size as 24 (months). Feature dimension of Google Trends data then can be expanded to 29*6*24=4176. The form of each Google Trends feature is in "keyword_category_n" ($1 \leq n \leq 24$).

3.3 Feature Selection

After data collection and preprocessing, each sample contains 4176 Google Trends features and 24 HPI features. This large feature scale has never been achieved by studies on house price prediction or Google Trends-related research. Such a large feature scale can obviously improve the performance of HPI growth rate prediction and find more interesting conclusions. However, some features might be redundant [31,32]. They will not only slow down the calculation process but also lead to overfitting and lower the modeling performance. Therefore, feature selection should be conducted in this study to eliminate the noise features and improve prediction performance.

This study adopted the RFE method for feature selection. GBDT is selected as the rank model of RFE. It can be seen from Table 1 that the number of features to remove per loop and the number of features to remain are two important parameters of RFE. In this paper, the former is pre-set as 1 and the latter needs to be further optimized. R-square is used to evaluate the performance of the models with different numbers of remained features. Higher value of R-square means better model performance.

To explore the effectiveness of Google Trends data for house price prediction, this study constructs three different models with different inputs: (1) $Model_{HPI}$ with HPI growth rate data only, (2) $Model_{Google\ Trends}$ with Google Trends data only, and (3) $Model_{HPI\ and\ Google\ Trends}$ with HPI growth rate data and Google Trends data. Five-fold cross-validation is applied to improve the robustness of the results. Time span h is set as 0, which indicates data of the previous 24 months are used to predict the HPI growth rate in the next month. The number of remained features in $Model_{Google\ Trends}$ is optimized in Figure 3 for an illustration. The results shows that the R-square value increases dramatically at first and then decreases slightly as the number of remained features increases. When the number of remained features is 59, the model obtains the highest R-square value. Therefore, the optimal number of remained features is 59 for $Model_{Google\ Trends}$. Similarly, the optimal number of remained features is 33 for $Model_{HPI\ and\ Google\ Trends}$ and 14 for $Model_{HPI}$.

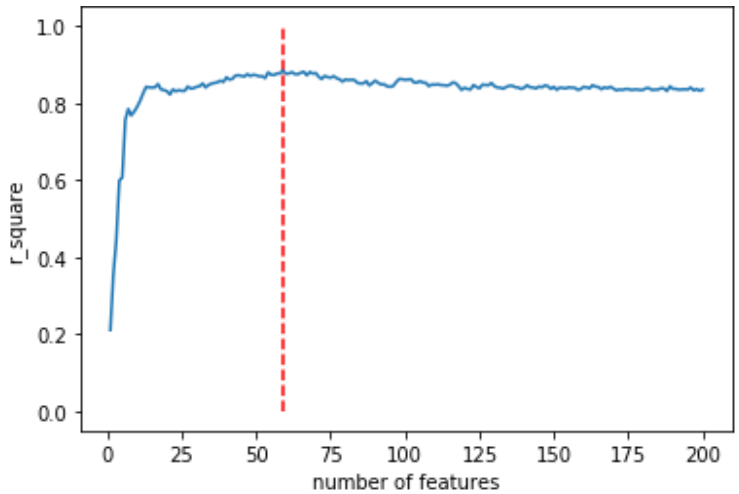


Figure 3. Optimization of the number of remained features in $Model_{Google\ Trends}$

The selected features for $Model_{Google\ Trends}$, $Model_{HPI\ and\ Google\ Trends}$ and $Model_{HPI}$ are presented in Table 3, Table 4 and Table 5, respectively. To further analyze the selected features, the category percentages of the selected features for $Model_{Google\ Trends}$ and $Model_{HPI\ and\ Google\ Trends}$ are presented in Table 6 (statistics from Table 3 and Table 4). The results show that 36% and 42% of the remained features are under "all categories" for $Model_{Google\ Trends}$ and $Model_{HPI\ and\ Google\ Trends}$ separately. The percentages of other categories for both models are similar (lower than 15%). This suggests that key words under "all categories" are closely related to HPI growth rate and can better reflect the search behavior of users.

Table 3. Selected features for $Model_{Google\ Trends}$

# feature	key word	category	lag
1	bank	All categories	20
2	bank	Real Estate	4
3	cpi	Business & Industrial	3

4	economic	All categories	15
5	economic	All categories	7
6	economic	Jobs & Education	14
7	economic crisis	All categories	19
8	economic growth	All categories	15
9	economic growth	All categories	3
10	economic growth	Jobs & Education	8
11	employment	Real Estate	23
12	financial crisis	All categories	8
13	financial crisis	Law & Government	22
14	gdp	Jobs & Education	9
15	hire	All categories	21
16	hire	All categories	22
17	hire	All categories	23
18	hire	All categories	24
19	hire	Business & Industrial	13
20	hire	Business & Industrial	24
21	house rent	All categories	11
22	house rent	All categories	12
23	house rent	Law & Government	10
24	housing market	All categories	16
25	housing market	All categories	23
26	housing market	Real Estate	13
27	immigration	All categories	8
28	immigration	All categories	9
29	immigration	Finance	15
30	immigration	Finance	7
31	immigration	Law & Government	17
32	immigration	Law & Government	19
33	income	All categories	13
34	income	All categories	14
35	income	All categories	15
36	income	Finance	14
37	income	Finance	2
38	income	Finance	22
39	income	Law & Government	14
40	invest	Jobs & Education	6
41	job	Business & Industrial	16
42	job	Business & Industrial	17
43	job	Finance	14
44	job	Finance	7
45	loan	Finance	17
46	loan	Jobs & Education	16
47	loan	Jobs & Education	17
48	loan	Jobs & Education	21
49	loan	Jobs & Education	22
50	loan	Real Estate	1
51	mortgage	Law & Government	3
52	mortgage	Law & Government	5
53	mortgage	Real Estate	18
54	real estate market	Real Estate	17
55	real estate market	Real Estate	18
56	salary	Real Estate	10
57	salary	Real Estate	11
58	stock	All categories	22
59	unemployment	Jobs & Education	24

Table 4. Selected features for $Model_{HPI\ and\ Google\ Trends}$

# feature	key word (or HPI growth rate)	category	lag
1	economic	All categories	22
2	economic	All categories	9
3	economic crisis	Business & Industrial	18
4	financial crisis	All categories	1
5	financial crisis	All categories	8
6	hire	All categories	22
7	immigration	All categories	8
8	immigration	Law & Government	7
9	income	All categories	12
10	income	All categories	13
11	income	All categories	24
12	income	Finance	13
13	income	Finance	2
14	income	Jobs & Education	9
15	income	Law & Government	12
16	income	Law & Government	13
17	HPI growth rate	/	1
18	HPI growth rate	/	10
19	invest	All categories	12
20	invest	Jobs & Education	5
21	job	Business & Industrial	17
22	land value	Finance	12
23	land value	Real Estate	5
24	loan	Business & Industrial	3
25	loan	Finance	16
26	loan	Jobs & Education	20
27	mortgage	Jobs & Education	16
28	mortgage	Real Estate	16
29	mortgage	Real Estate	18
30	real estate agency	All categories	5
31	real estate market	All categories	17
32	real estate market	All categories	18
33	salary	Real Estate	20

Table 5. Selected features for $Model_{HPI}$

window size	lag
24	1, 2, 3, 5, 6, 10, 11, 12, 15, 16, 18, 19, 22, 24

Table 6. Category percentage of the selected features for $Model_{Google\ Trends}$ and $Model_{HPI\ and\ Google\ Trends}$

category	$Model_{Google\ Trends}$	$Model_{HPI\ and\ Google\ Trends}$
All categories	36%	42%
Business & Industrial	8%	10%
Finance	14%	13%
Jobs & Education	15%	13%
Law & Government	12%	10%
Real Estate	15%	13%

4. Prediction Results and Analysis

4.1 Model Comparison

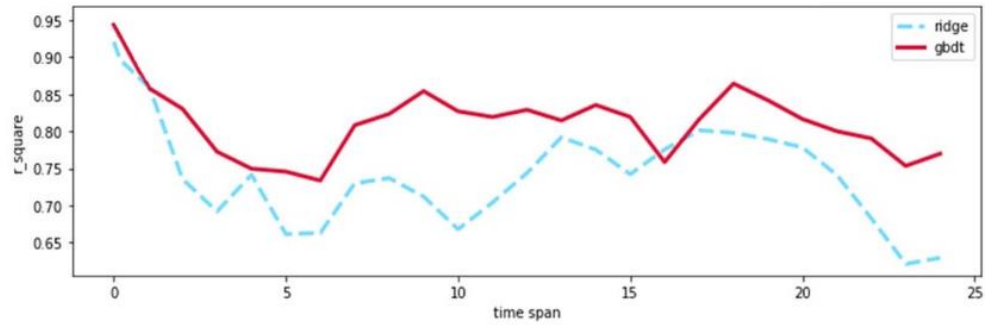
This paper aims at predicting the MoM growth rate of HPI using Google Trends data. GBDT algorithm is selected as the prediction model. Five-fold cross-validation is applied to improve the robustness of the results. To prove

that GBDT is a reasonable choice for house price prediction, Multiple Linear Regression (MLR), Ridge Regression, Support Vector Machine (SVM) and Random Forest (RF) are adopted in this paper as benchmarks. The comparison results of $Model_{HPI\ and\ Google\ Trends}$ are presented in Table 7. The results show that GBDT has better prediction performance than other methods on HPI growth rate prediction. GBDT obtains the highest R-square values of 0.944 in $Model_{HPI\ and\ Google\ Trends}$.

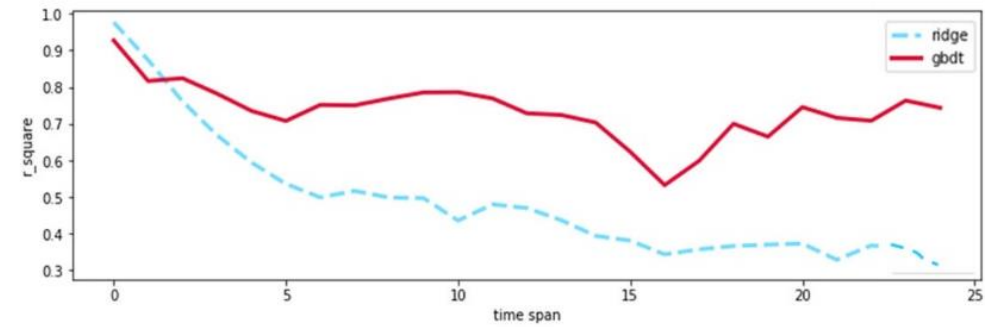
Table 7. Model comparison

Model	MLR	Ridge	SVM	RF	GBDT
$Model_{HPI\ and\ Google\ Trends}$	0.863	0.924	0.908	0.917	0.944

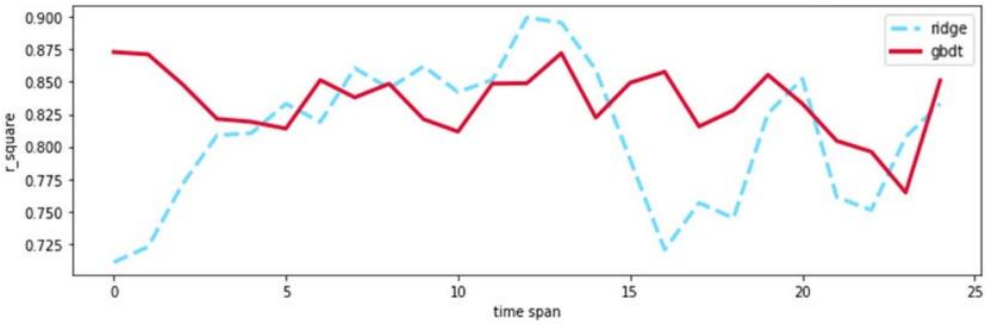
It is worth noting that results in Table 7 are given based on the time span of 0. As Ridge Regression and GBDT obtains the highest prediction performance in $Model_{HPI\ and\ Google\ Trends}$, to further analyze the modeling performance, prediction accuracy of these two methods for longer predicted time spans are compared. The range of time span is (0, 24). R-square values of the Ridge Regression and GBDT methods with different time spans are compared in Figure 4.



(a) $Model_{HPI\ and\ Google\ Trends}$



(b) $Model_{HPI}$



(c) $Model_{Google\ Trends}$

Figure 4. Comparison of Ridge Regression and GBDT using (a) HPI data and Google Trends data, (b) HPI data only, and (c) Google Trends data only

The results show that when the model uses HPI data and Google Trends data, or HPI data only as inputs, R-square values of GBDT overall are higher than Ridge Regression as the time span changes. Especially, as the time span becomes longer, R-square values of Ridge Regression keep decreasing while values of GBDT can keep high and stable. When only the Google Trends data are used as indicators, R-square values of the two models do not have too much difference. However, as the time span changes, values of GBDT model with Google Trends data are more stable than Ridge Regression model. Overall, it can be concluded that compared with Ridge Regression and other methods, GBDT has higher and more stable prediction performance for long term prediction. GBDT is a reasonable choice for HPI growth rate prediction.

4.2 Effectiveness of Google Trends Data

In this paper, a number of real-estate-related Google search indices are collected as the indicators of HPI growth rate prediction. To investigate whether these data can help improve the prediction accuracy, R-square values of GBDT models with different inputs are presented in Figure 5. The green dotted line represents the R-square values of $Model_{HPI\ and\ Google\ Trends}$, the black dotted line represents the R-square values of $Model_{HPI}$, and the red line represents the R-square of $Model_{Google\ Trends}$. The results show that R-square values of $Model_{HPI\ and\ Google\ Trends}$ and $Model_{Google\ Trends}$ are much higher and more stable than that of $Model_{HPI}$ as the time span becomes longer. This suggests that Google Trends data can effectively improve the performance of house price prediction, especially when the predicted time span is long. This is because Google Trends data can provide an insight into the housing market, and serve as a leading sentiment indicator of people's attitudes and expectations toward the house price. Google Trends indices are early market indicators and they can improve the prediction performance with long time span.

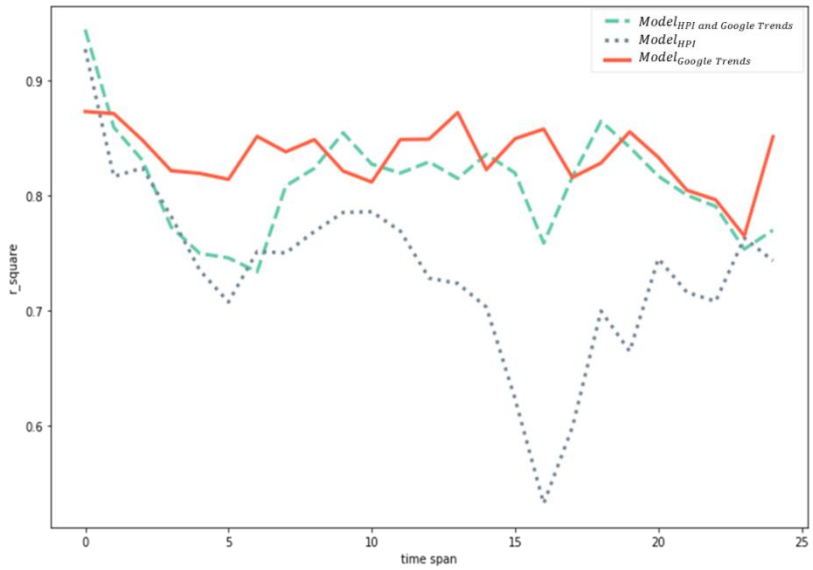


Figure 5. R-square values of GBDT with different model inputs

Table 8. Feature importance

Model inputs	Feature	Category	Lag	Average feature importance
Google Trends data only	Mortgage	Real estate	18	0.0898
	Housing market	All categories	23	0.0840
	Real estate agency	All categories	12	0.0770
	Real estate market	All categories	2	0.0475
	House rent	All categories	11	0.0378
Google Trends data and HPI data	HPI growth rate	\	1	0.2108
	Mortgage	Real estate	18	0.1234
	Real estate agency	All categories	5	0.0723
	Real estate market	All categories	17	0.0688
	House rent	All categories	2	0.0530

4.3 Feature Importance and Analysis

This paper has concluded that Google Trends index has considerable impacts on the prediction of HPI growth rate. To identify the important Google Trends features, two GBDT models are used for feature selection. One GBDT model only uses the Google Trends data as inputs and the other uses the HPI data and the Google Trends data as inputs. Time span for the predicted HPI growth rate is set as 0, 6, 12 and 24 months. Final feature importance is calculated based on the average feature importance of the four time spans. Top 5 most important features given by the two models are presented in Table 8.

The results show that the Top 5 most important features in two GBDT models are similar. To further understand the correlations between HPI growth rate and real-estate-related Google search indices, three categories of key words, including house rent, housing market & real estate market, and mortgage & real estate agency are analyzed in detail in the following of the study.

4.3.1 House Rent

To analyze the relationship between the Google Trends index of "house rent" and the MoM growth rate of HPI, variation of the search popularity of "house rent_all categories" during the study period is presented in Figure 6.

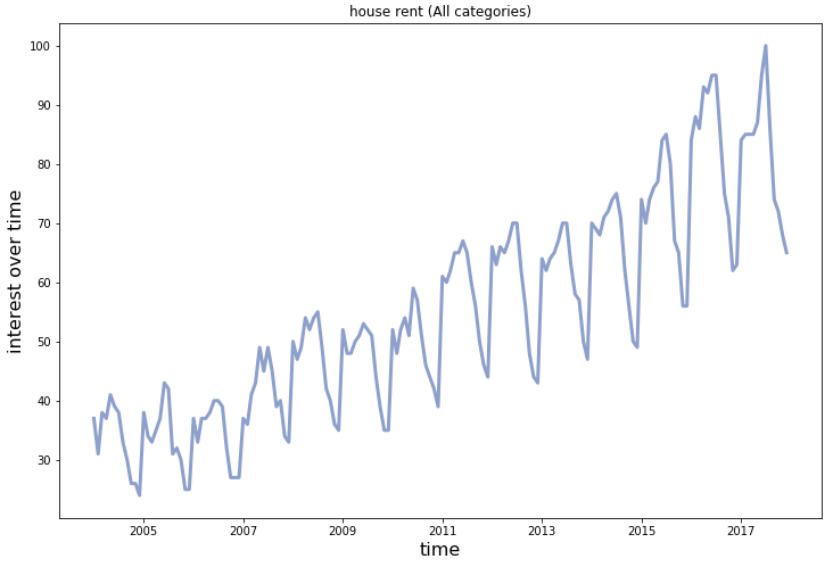


Figure 6. Variation of the search popularity of "house rent_all categories" during 2004 and 2017

The results show that people's interests on house rent keep increasing during 2004-2017. This suggests that people's needs for house renting are increasing. Figure 6 also shows that the variation of house rent follows a seasonal fluctuation. Its search popularity is relatively higher in the middle, or the summer of a year and lower in the winter of the year. This indicates that people's house renting needs are higher in summer. Possible reasons include: it would be more challenging to move in winter since most of cities in America have snowy or icy weather in winter; there are more holidays such as Easter and Christmas in winter and people might not want to be busy running for house moving during holidays; families with school-age children might be more like to move during the summer break [33].

Based on the same reasons, house buying might have the same seasonal fluctuation with house renting [33]. The demand for buying a house could be higher in summer but lower in winter. Higher demand, consequently, would lead to higher price [34]. Therefore, compared Figure 6 with Figure 2, it could be concluded that the seasonal fluctuation of the MoM growth rate of HPI might be partly caused by the seasonal demand for renting or buying houses.

4.3.2 Housing Market & Real Estate Market

"Housing market_all categories" and "real estate market_all categories" are another two important Google Trends indices related to HPI growth rate uncovered in this paper. Both of them reflect people's interests in buying, renting or investing in a house. Variations of the search popularity of these two features are presented in Figure 7.

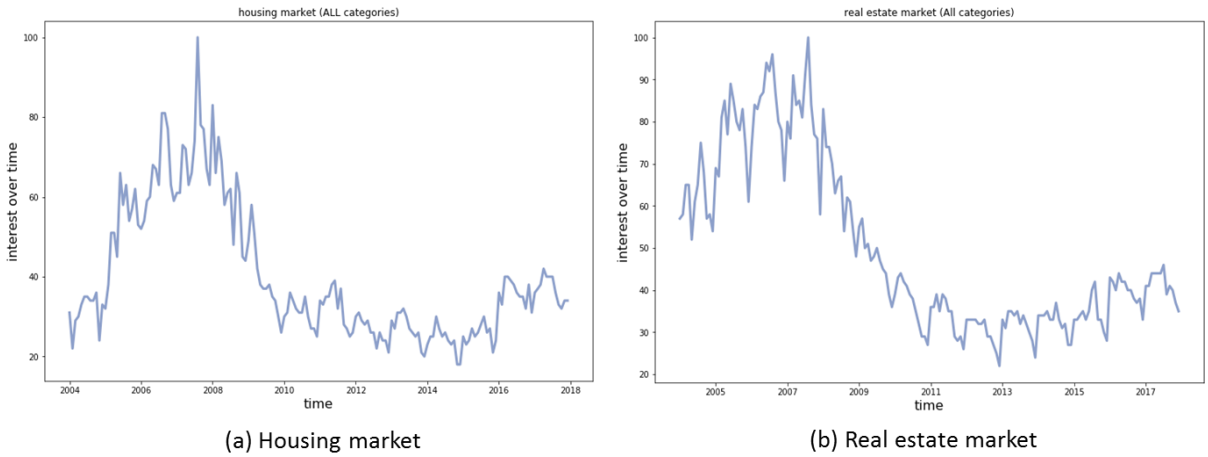


Figure 7. Variation of the search popularity of (a) housing market_all categories and (b) real estate market_all categories during 2004 and 2017

The results show that the search popularity of the two key words keeps increasing from 2004 and reaches the highest points around 2007 and 2008. This is because that in 2004, the America government increased the federal fund rate [35]. Consequently, the housing mortgage cost increased and the housing market bubble which had been accumulated since 2001 started to burst. People who has bought or invested in a house apparently wanted to know if their asset values would increase or decrease under the situation. Therefore, they would search more housing market or real estate market-related information on the Internet.

Compared Figure 7 with Figure 2, it can be seen that before the HPI growth rate reaches in lowest point in 2008, people's interests on housing market or real estate market have kept increasing in an abnormally fast speed. This suggests that people's abnormal high interests in housing market could be used as an indicator for an upcoming economic risk [36]. After 2008, the search popularity of the two key words decreases and gradually becomes stable. At the same time, HPI growth rate shows a slight increasing and stable changing trend. This further indicates that people's interest in housing market could be a useful indicator of the stability and health of the housing market.

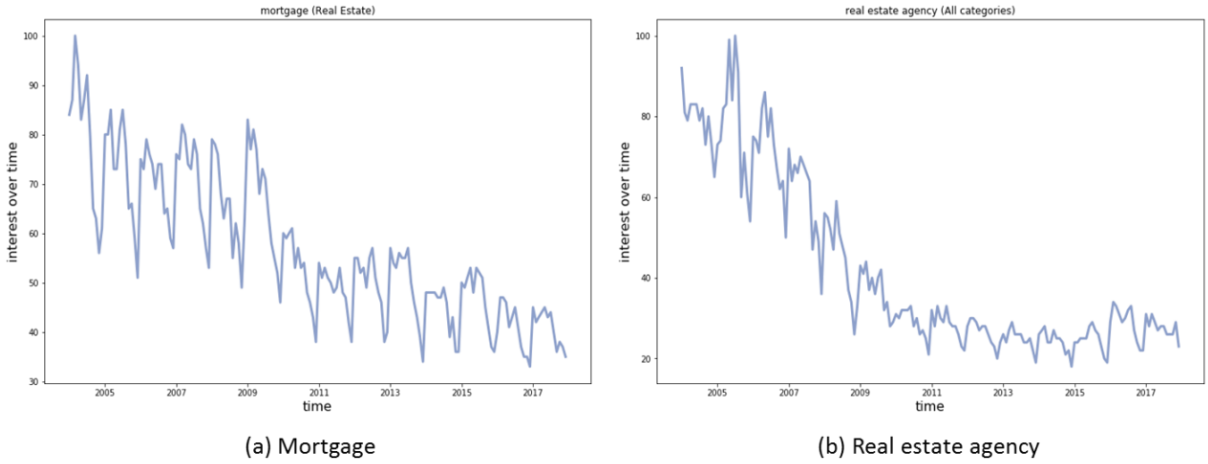


Figure 8. Variation of the search popularity of (a) mortgage_all categories and (b) real estate agency_all categories during 2004 and 2017

4.3.3 Mortgage & Real Estate Agency

"Mortgage_real estate" and "real estate agency_all categories" are two important factors people need before buying or investing in a house. Variations of the search popularity of these two features are shown in Figure 8. It can be seen that the search popularity of the two key words keeps decreasing in a fast speed from 2004 to 2010,

and then stays in a low and stable status. This is partly related to the increasing federal fund rate, too. As the cost of buying a house by mortgaging increases, people are less interested in finding a proper mortgage or a real estate agency to buy or invest in a house. As a result, the search popularity becomes low. Compared Figure 8 with Figure 2, it can be inferred that the decreasing demand of houses might be one of the reasons of the decreasing HPI growth rate from 2004 to 2008.

5. Conclusions

This study proposes a new methodology framework for house price prediction. Real estate-related Google Trends data, along with the fundamental HPI data, are collected to predict the MoM growth rate of HPI in the United States. A non-linear machine learning method namely GBDT is adopted as the major prediction model and the RFE model is utilized for feature selection. Three categories of models are constructed in this paper, including models with HPI data only, models with Google Trends data only, and models with HPI data and Google Trends data. The main conclusions are summarized as follows:

- RFE can effectively remove irrelevant or redundant features and improve the model performance.
- When selecting Google Trends data, the key words under the category of "all categories" usually can provide more information and are more important for HPI growth rate prediction.
- Compared with other prediction models, GBDT has higher and more stable performance for HPI growth rate prediction, especially when the predicted time span is long.
- Compared with models including fundamental HPI data only, models containing Google Trends data can exhibit higher and more stable prediction accuracy for long time span prediction.
- Three categories of Google Trends indices are the most important indicators of the HPI growth rate prediction. They are house rent, housing market & real estate market, and mortgage & real estate agency.

However, due to data availability, only the MoM growth rate of HPI is studied in this paper. Future work needs to collect more data for long-term HPI prediction (e.g., season-over-season or year-over-year HPI prediction) by the proposed framework and provide more insight views of HPI variation.

Reference

[1] Jiang, F., Ma, J., Webster, C. J., Chen, W., & Wang, W. (2024). Estimating and explaining regional land value distribution using attention-enhanced deep generative models. *Computers in Industry, 159–160*, Article 104103. https://doi.org/10.1016/j.compind.2024.104103

[2] Campbell, J. Y., & Cocco, J. F. (2007). How do house prices affect consumption? Evidence from micro data. *Journal of Monetary Economics, 54*(3), 591–621. https://doi.org/10.1016/j.jmoneco.2005.10.016

[3] Boelhouwer, P., Haffner, M., Neuteboom, P., & Vries, P. (2004). House prices and income tax in the Netherlands: An international perspective. *Housing Studies, 19*(3), 415–432. https://doi.org/10.1080/0267303042000204304

[4] Aoki, K., Proudman, J., & Vlieghe, G. (2004). House prices, consumption, and monetary policy: A financial accelerator approach. *Journal of Financial Intermediation, 13*(4), 414–435. https://doi.org/10.1016/j.jfi.2004.06.003

[5] Jiang, F., Ma, J., Webster, C. J., Wang, W., & Cheng, J. C. P. (2024). Automated site planning using CAIN-GAN model. *Automation in Construction, 159*, Article 105286. https://doi.org/10.1016/j.autcon.2024.105286

[6] Jiang, F., Ma, J., Webster, C. J., Li, X., & Gan, V. J. L. (2023). Building layout generation using site-embedded GAN model. *Automation in Construction, 151*, Article 104888. https://doi.org/10.1016/j.autcon.2023.104888

[7] Kouwenberg, R. R. P., & Zwinkels, R. C. J. (2011). Chasing trends in the U.S. housing market. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1539475

[8] Tsatsaronis, K., & Zhu, H. (2004). What drives housing price dynamics: Cross-country evidence. *Social Science Research Network.* Retrieved from https://papers.ssrn.com/abstract=1968425

[9] Deng, Y., Gyourko, J., & Wu, J. (2012). Land and house price measurement in China. *National Bureau of Economic Research.* https://doi.org/10.3386/w18403

[10] Bork, L., & Møller, S. V. (2015). Forecasting house prices in the 50 states using Dynamic Model Averaging and Dynamic Model Selection. *International Journal of Forecasting, 31*(1), 63–78. https://doi.org/10.1016/j.ijforecast.2014.05.005

[11] Jiang, F., & Ma, J. (2025). Environmental justice in the 15-minute city: Assessing air pollution exposure inequalities through machine learning and spatial network analysis. *Smart Cities, 8*, 53.

https://doi.org/10.3390/smartcities8020053

[12] Zhou, J., Li, Z., Ma, J. J., & Jiang, F. (2020). Exploration of the hidden influential factors on crime activities: A big data approach. *IEEE Access, 8*, 141033–141045. https://doi.org/10.1109/ACCESS.2020.3009969

[13] Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018). House price prediction using machine learning and neural networks. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1936–1939). IEEE. https://doi.org/10.1109/ICICCT.2018.8473231

[14] Wei, Y., & Cao, Y. (2017). Forecasting house prices using dynamic model averaging approach: Evidence from China. *Economic Modelling, 61*, 147–155. https://doi.org/10.1016/j.econmod.2016.12.002

[15] Schäfers, W., Braun, N., & Dietzel, M. A. (2014). Sentiment-based commercial real estate forecasting with Google search volume data. *Journal of Property Investment & Finance, 32*(6), 540–569. https://doi.org/10.1108/JPIF-01-2014-0004

[16] Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record, 88*(s1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

[17] Jiang, F., Yuen, K. K. R., & Lee, E. W. M. (2020). Analysis of motorcycle accidents using association rule mining-based framework with parameter optimization and GIS technology. *Journal of Safety Research, 75*, 292–309. https://doi.org/10.1016/j.jsr.2020.09.004

[18] Jiang, F., Yuen, K. K. R., Lee, E. W. M., & Ma, J. (2020). Analysis of run-off-road accidents by association rule mining and geographic information system techniques on imbalanced datasets. *Sustainability, 12*(12), Article 4882. https://doi.org/10.3390/su12124882

[19] Li, Z., Ma, J., & Jiang, F. (2024). Exploring the effects of 2D/3D building factors on urban energy consumption using explainable machine learning. *Journal of Building Engineering, 97*, Article 110827. https://doi.org/10.1016/j.jobe.2024.110827

[20] Jiang, F., Ma, J., Webster, C. J., Chiaradia, A. J. F., Zhou, Y., Zhao, Z., & Zhang, X. (2024). Generative urban design: A systematic review on problem formulation, design generation, and decision-making. *Progress in Planning, 180*, Article 100795. https://doi.org/10.1016/j.progress.2023.100795

[21] Breiman, L. (2017). Classification and regression trees. *Routledge*. https://doi.org/10.1201/9781315139470

[22] Si, S., Zhang, H., Keerthi, S. S., Mahajan, D., Dhillon, I. S., & Hsieh, C.-J. (n.d.). Gradient boosted decision trees for high dimensional sparse output.

[23] Jiang, F., Ma, J., & Li, Z. (2022). Pedestrian volume prediction with high spatiotemporal granularity in urban areas by the enhanced learning model. *Sustainable Cities and Society, 79*, Article 103653. https://doi.org/10.1016/j.scs.2021.103653

[24] Mohan, A., Chen, Z., & Weinberger, K. (n.d.). Web-search ranking with initialized gradient boosted regression trees.

[25] Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems, 83*(2), 83–90. https://doi.org/10.1016/j.chemolab.2006.01.007

[26] Jiang, F., & Ma, J. (2021). A comprehensive study of macro factors related to traffic fatality rates by XGBoost-based model and GIS techniques. *Accident Analysis & Prevention, 163*, Article 106431. https://doi.org/10.1016/j.aap.2021.106431

[27] Jiang, F., Yuen, K. K. R., Lee, E. W. M., & Ma, J. (2020). A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. *Accident Analysis & Prevention, 141*, Article 105520. https://doi.org/10.1016/j.aap.2020.105520

[28] Freddie Mac Home. (n.d.). http://www.freddiemac.com//index.html

[29] Google Trends. (2019). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Google_Trends&oldid=883779596

[30] Time series analysis. (n.d.). Princeton University Press. https://press.princeton.edu/titles/5386.html

[31] Jiang, F., & Ma, J. (2025). Predicting urban vitality at regional scales: A deep learning approach to modelling population density and pedestrian flows. *Smart Cities, 8*, 58. https://doi.org/10.3390/smartcities8020058

[32] Jiang, F., Ma, J., Li, Z., & Ding, Y. (2022). Prediction of energy use intensity of urban buildings using the

semi-supervised deep learning model. *Energy, 249*, Article 123631. https://doi.org/10.1016/j.energy.2022.123631

[33] Kajuth, F. (n.d.). Seasonality in house prices.

[34] Green, R., & Hendershott, P. H. (1996). Age, housing demand, and real house prices. *Regional Science and Urban Economics, 26*(5), 465–480. https://doi.org/10.1016/0166-0462(96)02128-X

[35] Federal funds rate. (2019). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Federal_funds_rate&oldid=881182406

[36] Dietzel, M. A. (2016). Sentiment-based predictions of housing market turning points with Google trends. *International Journal of Housing Markets and Analysis, 9*(1), 108–136. https://doi.org/10.1108/IJHMA-12-2014-0058