Beyond GWR: A Spatial Clustering and Partitioned Stacking Framework for Capturing Nonlinear Locational Premiums and Submarket Specificity in Real Estate Valuation

Zezhong Wang , Wanxin Li , Xiaolin Sun , Shuohan Jiang , Jing Li *

Posted Date: 3 March 2026

doi: 10.20944/preprints202603.0178.v1

Keywords: spatial clustering; stacking ensemble learning; housing price prediction; locational premium; submarket heterogeneity

 doi:10.20944/preprints202603.0178.v1

Article

Beyond GWR: A Spatial Clustering and Partitioned Stacking Framework for Capturing Nonlinear Locational Premiums and Submarket Specificity in Real Estate Valuation

Zezhong Wang [1], Wanxin Li [2], Xiaolin Sun [1],Shuohan Jiang [3] and Jing Li [4,*]

1. College of Civil and Architectural engineering, North China University of Science and Technology, Tangshan Hebei 063210, China
2. College of Sciences, North China University of Science and Technology, Tangshan Hebei 063210, China
3. School of Economics and Management, North China University of Science and Technology, Tangshan Hebei 063210, China
4. College of Mining Engineering, North China University of Science and Technology, Tangshan Hebei 063210, China

* Correspondence: ljing@ncst.edu.cn; Tel.:+86-150-8192-3700

Abstract

Based on a spatial clustering and partitioned stacking ensemble model, this study addresses the limitations of traditional geoweighting regression in capturing nonlinear location premiums and submarket heterogeneity within urban real estate markets. It proposes a two—stage modeling framework: "spatial clustering → partitioned differentiated stacking ensemble." Using long—term multi—source transaction data for Beijing's secondary housing market, the study divides the market into three spatially heterogeneous submarkets: core, near—suburban, and far—suburban. Stacked ensemble models based on ElasticNet, XGBoost, LightGBM, and Random Forest are constructed within each submarket. Factor analysis extracts interpretable common factors, which are combined with Lasso and SHAP for feature selection and impact mechanism analysis. Results indicate that the zoned stacking model performs exceptionally well across all three submarkets, achieving an R^2 of 0.916 in the core urban area. Significant nonlinear location premiums exist within the core urban area. The multi—level interpretability framework reveals the differentiated effects of location and scale factors across different submarkets. This study advances from "global modeling" to "spatial zoning + adaptive ensemble," providing a viable tool for refined valuation and risk management in highly heterogeneous markets.

Keywords: spatial clustering; stacking ensemble learning; housing price prediction; locational premium; submarket heterogeneity

1. Introduction

The accurate valuation of second—hand housing remains a persistent and complex challenge, primarily because of the inherent spatial heterogeneity, information asymmetry, and multifaceted, nonlinear price drivers that are characteristic of urban real estate markets[1]. As the global economy increasingly depends on real estate as a cornerstone of economic stability and growth, and as markets such as China make a decisive transition from an "incremental" to a "stock"-driven paradigm, the demand for robust, adaptable, and interpretable valuation frameworks has never been greater[2,3].This requires moving beyond traditional statistical methods and geographically weighted regression (GWR) approaches, which often have difficulty fully capturing local complexities and modeling interactions[4,5]. Instead, the field should embrace more sophisticated

 doi:10.20944/preprints202603.0178.v1

data—driven methodologies that can effectively integrate spatial intelligence with the predictive power of modern ensemble learning techniques. This study addresses this need by proposing a novel Spatial Clustering and Partitioned Stacking Framework, aiming to push forward the methodological frontier in real estate valuation by better capturing nonlinear locational premiums and submarket—specific price formation mechanisms

1.1. The Economic Significance and Transition of the Chinese Housing Market

The real estate sector is a significant pillar of national economic development and stability, making substantial contributions to GDP and employment [6]. Particularly, China's housing market is undergoing a profound structural transformation from a development—centered "incremental era" to a transaction—oriented "stock era" [3].

In this new model, the volume and economic significance of second—hand housing transactions have exceeded those of new housing in many major cities, making the secondary market the core of the residential real estate sector. However, this market is characterized by pronounced spatial heterogeneity (where housing attributes and values vary significantly across different urban districts), severe information asymmetry (between buyers, sellers, and intermediaries), and high policy sensitivity (to regulations on purchase restrictions, loans, and taxation) [7–10].

These characteristics jointly pose challenges to traditional appraisal methods and hedonic pricing models, which often assume spatial stationarity or rely on simplified linear relationships. Consequently, there is an urgent need to develop advanced, spatially—aware predictive frameworks that integrate machine learning to better interpret complex price dynamics, support evidence—based decision—making for various stakeholders (including homeowners, investors, financial institutions, and policymakers), and enhance market transparency and efficiency in multi—tiered urban contexts.

1.2. Evolution of Predictive Modeling: From Traditional Methods to Machine Learning

The quest for accurate real estate valuation has driven the evolution of predictive modeling techniques. Early scholarly and practical efforts heavily relied on traditional statistical models. Time—series approaches, such as Autoregressive (AR) models [11–13] and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models [14,15], were employed to capture temporal trends, seasonality, and volatility clustering in price indices. Cross—sectional methods, notably hedonic price models and their spatial extensions like Geographically Weighted Regression (GWR), attempted to attribute value to individual property characteristics and location.

While useful for identifying broad trends and average relationships, these conventional methods often fall short in handling the inherent nonlinearity, complex interaction effects, and high—dimensional, heterogeneous nature of data in modern housing markets. Their reliance on parametric assumptions and limited capacity to model intricate patterns can lead to suboptimal predictive accuracy and poor generalization [16,17].

The advent of machine learning (ML) has ushered in a new era for real estate analytics. Algorithms such as Random Forests (RF) [18], eXtreme Gradient Boosting (XGBoost) [19,20], Gradient Boosting Trees (GBT) [21], and various Neural Network (NN) architectures [22,23] have demonstrated superior capability in modeling complex, nonlinear relationships without stringent prior assumptions. These data—driven methods excel at learning intricate patterns from large, multifaceted datasets encompassing structural attributes, locational features, market sentiments, and macroeconomic indicators.

Comparative studies have consistently validated the predictive accuracy advantages of these ensemble and deep learning models over traditional econometric techniques [24,25]. This evolution underscores a critical shift towards leveraging computational power and algorithmic sophistication to enhance valuation precision.

 doi:10.20944/preprints202603.0178.v1

1.3. Research Gaps and the Potential of Stacking Fusion Frameworks

Despite the promising advances brought by individual ML algorithms, significant research gaps remain in the application of these techniques to second—hand housing valuation.

First, there is a predominant focus on applying and tuning single—algorithm models (e.g., optimizing an XGBoost model in isolation). This approach neglects the potential of model fusion strategies, which can synergistically combine the strengths of diverse learners to achieve more robust and accurate predictions than any single model.

Second, the handling of high—dimensional feature spaces—common in real estate due to the multitude of potential price drivers—is often rudimentary. Studies frequently resort to simple correlation—based filtering or principal component analysis (PCA) for dimensionality reduction, which may discard valuable predictive information. There is a lack of work that strategically combines factor analysis (to extract latent constructs) with regularization techniques like LASSO (to perform feature selection within the derived factors) for a more nuanced and powerful feature engineering pipeline.

Third, a persistent trade—off between model interpretability and predictive power exists. While complex ensemble or neural models may achieve high accuracy, they often function as "black boxes," limiting their practical utility for stakeholders who require insights into why a property is valued a certain way. This gap restricts the models' applicability in scenarios requiring explainable decisions, such as mortgage underwriting or policy impact assessment.

These limitations highlight the need for a more integrated, sophisticated modeling paradigm. Stacking (or Stacked Generalization) fusion frameworks present a compelling solution to address these gaps. Unlike simple averaging or voting ensembles, stacking employs a meta—learner to optimally combine the predictions of multiple, diverse base models (e.g., linear models, tree—based models). This architecture has demonstrated remarkable success and generalization capability across diverse fields beyond real estate, including electromagnetic effect prediction [26], pavement roughness forecasting [27], gas outburst risk warning [28], and intelligent transportation systems [29,30]. The core strength of stacking lies in its ability to leverage the unique inductive biases of different base models, allowing the meta—learner to correct individual model errors and capture patterns that may be missed by any single algorithm.

This study posits that the full potential of stacking for real estate valuation can only be unlocked when it is integrated with a spatial segmentation strategy. The pronounced spatial heterogeneity of housing markets means that a single, global model is likely to be suboptimal. Therefore, we propose a novel Spatial Clustering and Partitioned Stacking Framework. The framework first uses spatial clustering to partition the urban market into geographically and economically homogeneous submarkets. Then, for each identified submarket, a dedicated stacking ensemble model is trained. This two—stage approach—"clustering then stacking"—is designed to move decisively "Beyond GWR." While GWR allows parameters to vary across space, it remains within the linear or generalized linear modeling family. Our framework, by contrast, enables the application of highly flexible, non—linear ensemble methods that are explicitly tailored to the distinct price formation mechanisms of each submarket. This allows for the precise capture of non—linear locational premiums (e.g., threshold effects near amenities) and submarket specificity, ultimately aiming to deliver superior predictive accuracy alongside enhanced interpretability through submarket—level analysis.

The remainder of this paper details the methodology, presents an empirical application to Beijing's second—hand housing market, discusses the results, and concludes with implications for research and practice.

 doi:10.20944/preprints202603.0178.v1

2. Materials and Methods

2.1. Research Objective and Overall Framework

This study proposes a data-driven framework for accurately predicting transaction prices of second-hand housing in Beijing. The framework is structured into three sequential stages, as illustrated in the overall workflow (Figure 1). First, spatial clustering is applied to partition the housing market into distinct geographic submarkets based on their inherent price formation mechanisms. Second, differentiated modeling is conducted for each identified submarket, wherein tailored data preprocessing and predictive modeling pipelines are developed. Within each submarket, a stacked ensemble learning architecture is employed, integrating the predictive capabilities of multiple base learners—ElasticNet, XGBoost, LightGBM, and Random Forest. Third, an integrated meta-modeling stage compares the performance of linear regression and decision tree as meta-learners to combine the base learners' predictions optimally. The final output of the framework delivers not only precise transaction price forecasts but also explainable insights into the key drivers of housing prices across different submarkets.

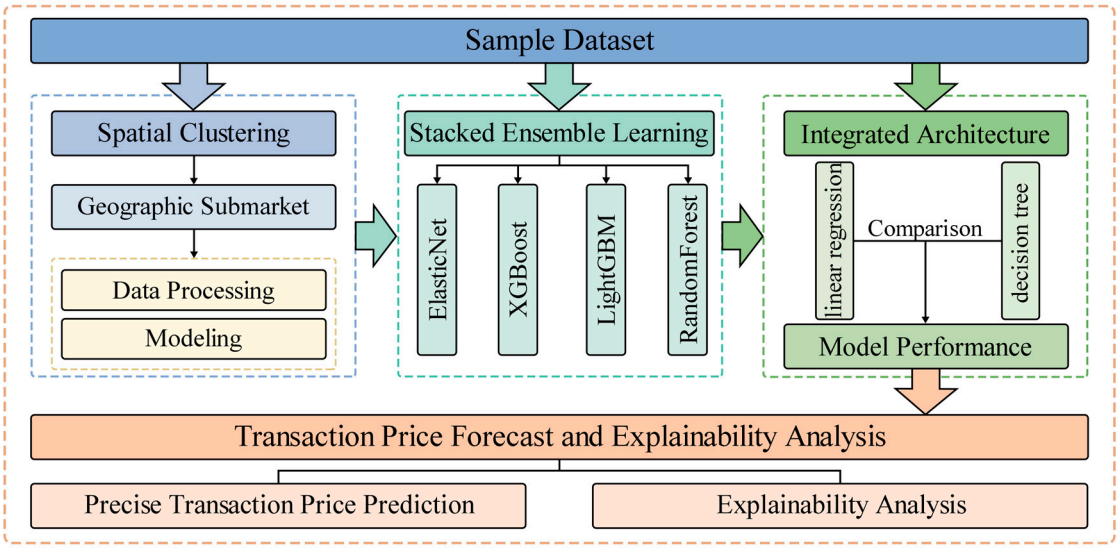


Figure 1. Overall Framework.

2.2. Data Preprocessing

The data underwent the following preprocessing steps:

- Data Cleaning: Records with extreme outliers or significant missing values in core variables (e.g., price and area) were removed.
- Feature Engineering: Spatial features were derived from geographic coordinates, and continuous variables were standardized to eliminate scale effects.
- Feature Dimensionality Reduction: Factor analysis was applied to address multicollinearity and extract interpretable common factors as key input variables for subsequent modeling.

The factor analysis model is defined as follows[31,32] (1):

$$X = (X_1, X_2, \cdots X_p) \quad (1)$$

It was a p-dimensional column vector consisting of the original observed variables, and let$F_1, F_2, \ldots F_m$be an m-dimensional column vector composed of the newly derived common factors, where these m common factors are mutually uncorrelated.

In factor analysis, a slight "Heywood case" may occasionally occur when the model approaches the fitting limit of the data or when the sample size is relatively limited. This phenomenon is characterized by the estimated variance of one or more common factors slightly exceeding the

 doi:10.20944/preprints202603.0178.v1

theoretical range of 0 to 1, and it is generally regarded as a boundary solution. In this study, under the trade-off between maximizing data reduction and preserving predictive utility, this situation was encountered. The usability of the model was comprehensively evaluated by examining its overall goodness-of-fit and predictive validity.

 doi:10.20944/preprints202603.0178.v1

2.3. Lasso-Based Feature Selection

Feature selection aims to identify a subset of the most informative features. Common approaches are categorized into three types: filter methods, wrapper methods, and embedded methods[33]. Although prior dimensionality reduction via factor analysis and subsequent orthogonal rotation yielded a set of uncorrelated factor vectors — effectively filtering out noise and redundant information—further feature selection was deemed necessary to refine the predictor set. To this end, Lasso (Least Absolute Shrinkage and Selection Operator) regression was employed[34]. Using the LassoCVfunction from the sklearnlibrary, hyperparameter tuning was performed via cross-validation to optimize the regularization strength. The Lasso method applies an L1 penalty to the regression coefficients, effectively shrinking some coefficients to zero and thereby performing automatic feature selection. Following this procedure, both a feature importance plot and a regularization path plot were generated. Based on a comprehensive evaluation of feature importance rankings and model performance metrics, a final set of retained features was determined for subsequent predictive modeling.

2.4. Model Construction Method

The methodological workflow consists of three main stages:

Stage 1: Spatial Clustering for Submarket Identification

To address the limitations of "global models" in capturing spatial heterogeneity, spatial clustering algorithms (e.g., DBSCAN, spatially constrained K—Means, or Spectral Clustering) were applied to the preprocessed data. This step divides Beijing's second—hand housing market into several spatially contiguous or neighboring submarkets (clusters), each characterized by relative internal homogeneity and distinct price—driving mechanisms.

Stage 2: Partition—Specific Modeling and Stacked Ensemble Learning

For each identified submarket, an independent modeling pipeline was implemented:

- Base Model Training: Four advanced machine learning algorithms with different inductive biases—ElasticNet, XGBoost, LightGBM, and RandomForest—were trained to capture the complex relationships between housing prices and feature factors within each submarket.
- Stacked Ensemble Learning: To combine the strengths of multiple base models and improve predictive robustness, a stacked ensemble approach was adopted. Predictions from the base models on the validation sets were used as meta—features and input into meta—models. The performance of linear regression (LR) and decision trees (DT) as meta—models was compared to assess the necessity of complex meta—model architectures.

Stage 3: Model Evaluation and Interpretability Analysis

Model performance was thoroughly evaluated using hold—out or cross—validation methods, with metrics including Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R—squared (R^2). To extract actionable insights, explainable AI techniques—specifically SHAP (SHapley Additive exPlanations)[35]—were employed to quantify the contribution of each feature factor in the final ensemble model, with particular emphasis on revealing the nonlinear influence patterns and interaction effects of key factors such as "location."

2.5. Model Training and Evaluation

A Bayesian Optimization framework with the Tree-structured Parzen Estimator (TPE) algorithm was employed to automatically search for the optimal hyperparameters of each base model[36,37]. The objective was to maximize the average performance measured via 5-fold cross-validation. During the training of the stacked ensemble model, out-of-fold (OOF) predictions generated from cross-validation were used as meta-features to prevent data leakage. Model performance was comprehensively evaluated using the following metrics:

- Adjusted R^2 in Equation (2): Measures the proportion of variance explained by the model while penalizing model complexity.

 doi:10.20944/preprints202603.0178.v1

- Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) in Equation (3)-(4): Quantify the magnitude of prediction errors in monetary terms (in units of 10,000 RMB per square meter).

Mean Absolute Percentage Error (MAPE) in Equation (5): Provides a percentage-based view of prediction error relative to the actual value, facilitating business interpretation.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(T_i - T_p)^2}{\sum_{i=1}^{n}(T_i - T_a)^2} \tag{2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n} |T_p - T_i| \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(T_p - T_i)^2}{n}} \tag{4}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \frac{|T_p - T_i|}{|T_i|} \tag{5}$$

2.6. Interpretability Analysis

To move beyond "black-box" predictions and validate the reasonableness of the factor analysis, this study applied the SHapley Additive exPlanations (SHAP) method for post-hoc interpretation of the best-performing model [38,39]. Grounded in cooperative game theory, SHAP assigns each feature a contribution value for every individual prediction. This approach enables both global identification of key driving factors and local explanation of individual predictions, thereby supporting the linkage between data-driven insights and real estate economics theory

3. Results and Discussion

3.1. Data Sources and Preprocessing

The dataset for this study consists of transaction records of Beijing's second—hand housing market from 2010 to 2024, aggregated from a leading real estate transaction platform. The original dataset, which contains over one million records, encompasses three primary dimensions:

1) Physical Property Attributes, including area, layout, floor level, orientation, construction year, and renovation status;

2) Spatial Location Information, such as latitude, longitude, and administrative district;

3) Market Dynamics Indicators, including listing price, number of interested buyers, page view counts, price adjustment history, and transaction cycle duration.

Given the extensive influence, frequent issuance, and complex scoring mechanisms of Beijing's housing market policies within short intervals, explicit policy factors were excluded from the model to maintain analytical clarity.

A systematic data cleaning and preprocessing pipeline was employed to ensure data quality. Initially, missing values were dealt with: numerical features were imputed using the median, categorical features using the mode, and features with over 30% missing data were removed. Subsequently, outliers in the target variable (transaction price) were managed using a threshold method based on the 99.5th percentile, supplemented with the Isolation Forest algorithm to alleviate the impact of extreme values on model training. Finally, all date fields were decomposed and encoded, while text—based fields (e.g., orientation, floor descriptions) were vectorized. All input features were then standardized to a consistent numerical format appropriate for model processing. This rigorous preprocessing produced a high—quality, coherent analytical dataset, providing a robust foundation for subsequent modeling stages.

 doi:10.20944/preprints202603.0178.v1

3.2. Feature Engineering

3.2.1. Spatial Clustering Analysis

To validate and quantify the spatial heterogeneity in Beijing's secondary housing market, this study employs unsupervised learning methods to partition the market. Using administrative districts as the basic unit, we calculate the statistical characteristics, including the annual average housing price and price variance, for each district. Subsequently, the K—Means++ clustering algorithm is applied to analyze these multidimensional features. Based on historical experience and practical information, the optimal number of clusters is determined to be 3. The calculated silhouette coefficient of 0.558 > 0.5 confirms that this clustering aligns with empirical assumptions and has a sound theoretical classification basis. Based on this, Beijing is divided into three sub—markets that are homogeneous internally but heterogeneous from each other (Cluster 1: Pink, 2: Blue, 3: Green), providing a foundation for subsequent zoning modeling.

This study employs the K—Means++ clustering algorithm to partition the 16 administrative districts of Beijing into three distinct housing sub—markets, with a silhouette coefficient of 0.558 indicating effective clustering. The results visually demonstrate a three—tier gradient structure of "core–suburban–exurban" spatial differentiation: the core area (Cluster 2) exhibits high value and high volatility, reflecting resource scarcity and internal value segmentation; the suburban area (Cluster 1) shows moderate value and volatility, indicating transitional characteristics driven by planning policies; and the exurban area (Cluster 0) displays low value and low volatility, representing a relatively homogeneous market dominated by basic residential functions. This spatial classification quantitatively confirms the regional heterogeneity in Beijing's second—hand housing market and provides a critical foundation for the subsequent implementation of differentiated, sub—market—specific modeling.

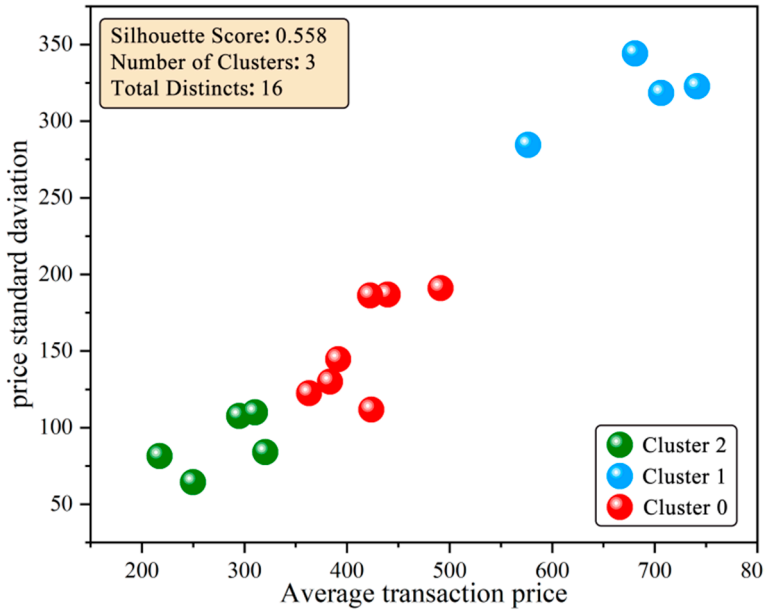


Figure 2. Clustering results based on average transaction price and price standard deviation using the K-Means++ algorithm.

Table 1 presents the geographic classification of Beijing's administrative districts into three distinct sub—markets (Clusters) based on spatial clustering analysis. Cluster 2 represents the core urban districts, Cluster 1 consists of suburban transitional areas, and Cluster 3 corresponds to outer suburban districts. This classification provides the spatial basis for the subsequent differentiated modeling of housing price determinants across sub—markets.

 doi:10.20944/preprints202603.0178.v1

Table 1. Spatial Clustering Results of Beijing's Second-hand Housing Market Based on K-Means++.

Cluster 1	Cluster 2	Cluster 3
Fengtai, Yizhuang Development Zone, Changping, Shijingshan, Tongzhou, Daxing, Shunyi	Xicheng, Dongcheng, Haidian, and Chaoyang districts	Fangshan, Mentougou, Huairou, Pinggu, Miyun

The plot compares the price distributions for Cluster 0 (red), Cluster 1 (blue), and Cluster 2 (green). The width of each violin represents the data density at given price levels, while the inner box—plots indicate the median, interquartile range, and potential outliers.The visualization clearly reflects a spatial price gradient: Cluster 2 (core urban districts) exhibits the highest and most dispersed prices, with a right—skewed distribution; Cluster 1 (suburban transitional areas) shows moderate price levels with relatively symmetric density; Cluster 0 (outer suburban districts) displays the lowest and most concentrated price distribution.This figure underscores the pronounced spatial heterogeneity in housing prices, supporting the necessity of sub—market—specific modeling approaches.

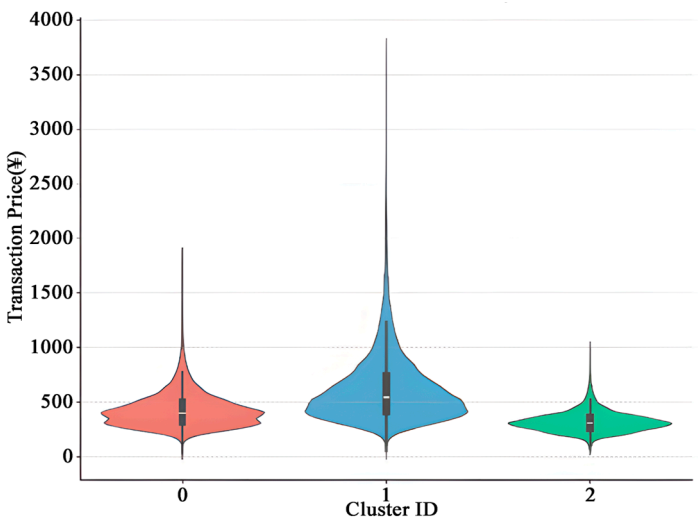


Figure 3. Violin plots with embedded boxplots illustrating the distribution of transaction prices across three clustered sub—markets in Beijing's second—hand housing market.

3.2.2. Dimensionality Reduction

This study first preprocessed the raw data to construct an initial feature set. To preliminarily explore the linear correlations among features, Pearson correlation coefficient heatmaps were plotted for three categories of features, as shown in Figure 4.

Subsequently, factor analysis was employed to reduce dimensionality and extract latent structures by constructing a factor model. Based on the correlation matrix, this method transforms the original multiple observed variables into a few comprehensive variables (i.e., common factors) with clear economic interpretations.

By analyzing the rotated factor loading matrix, each common factor was named according to the practical significance of variables with high loadings. Finally, the composite score of each sample on these factors was calculated and used as input features for subsequent modeling.

 doi:10.20944/preprints202603.0178.v1

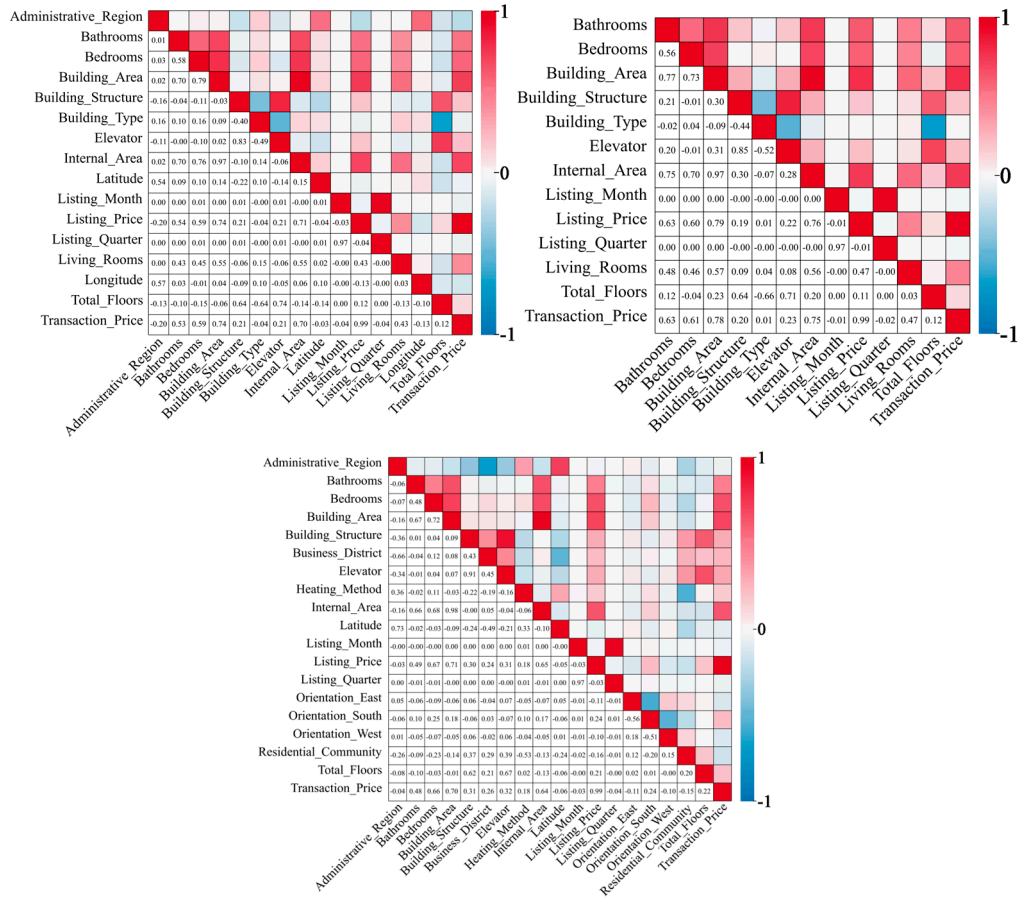


Figure 4. Pearson correlation heatmaps among property features grouped by three feature categories.

3.2.3. Variable Construction

To comprehensively capture the multidimensional factors influencing housing prices, this study conducted systematic feature engineering [40,41]. Specifically, composite structural features were decomposed: the unit layout was broken down into independent count variables for "rooms," "living rooms," "kitchens," and "bathrooms." Floor information was standardized and encoded into categories such as basement, low-rise, mid-rise, and high-rise, with a floor position ratio (current floor/total floors) also calculated. Orientation was transformed into four binary features indicating east, south, west, and north exposures. Area units were standardized, and a floor–unit ratio was computed. Additionally, building age (e.g., whether the property is over two or five years old) and floor category (high, middle, low) were encoded as categorical variables, whereas the exact floor number was retained as a numerical variable. Through the above transformations, a set of 36 initial features was constructed, as summarized in the following Table 2:

Table 2. Variable Construction.

type of variable	Feature Name
numeric type	Listing price, price adjustment (times), viewings (times), followers (people), views (times), net area, floor-to-ceiling ratio, gross floor area, longitude, latitude, listing year, listing quarter, listing month, rooms, living room, kitchen, bathroom, floor number, east, south, west, north
By typeBy type	administrative district, building type, house age, property ownership, floor level, unit layout, decoration status, building structure, heating method, elevator availability, transaction ownership, property use, business district, residential community

 doi:10.20944/preprints202603.0178.v1

3.2.4. Factor Analysis

To assess the suitability of the data for factor analysis(Table 3), the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy [42,43]and Bartlett's test of sphericity[44] were separately conducted for each of the three clustered submarkets using functions from the 'factor_analyzer' library ('calculate_kmo' and 'calculate_bartlett_sphericity').

The KMO values for all three submarkets exceeded 0.6 (Cluster 1: 0.7363, Cluster 2: 0.7317, Cluster 3: 0.6727), indicating that the partial correlations among variables are sufficiently strong and that the data are appropriate for factor analysis. Specifically, Clusters 1 and 2 exhibited higher KMO values (both > 0.73), suggesting more pronounced inter—variable correlations and potentially better factor extraction performance. Although Cluster 3's KMO value was comparatively lower, it remained within an acceptable range, implying somewhat weaker correlations while still meeting the basic requirements for factor analysis.

Bartlett's test yielded highly significant results for all three submarkets, with p—values approaching zero and large chi—square statistics (Cluster 1: 1,012,164; Cluster 2: 951,913; Cluster 3: 150,991). These results strongly reject the null hypothesis of variable independence, confirming the presence of significant correlational structures among the features within each submarket and providing statistical justification for proceeding with factor analysis.

In summary, both KMO and Bartlett's test results consistently indicate that the variables share sufficient common variance and exhibit significant correlations, making the data highly suitable for factor analysis. These findings lay a solid foundation for subsequent extraction of common factors, dimensionality reduction, and factor model construction. Furthermore, the observed variation in KMO values across submarkets preliminarily reflects differences in the strength of variable correlations, which may correspond to heterogeneity in the determinants of housing prices across different regional or market tiers—an aspect that warrants further attention during factor extraction and interpretation.

Table 3. Results of KMO and Bartlett's Tests.

Test	Cluster 1	Cluster 2	Cluster 3
Kaiser–Meyer–Olkin Measure	0.7363	0.7317	0.6727
Bartlett's Test of Sphericity (χ2)	1,012,164	951,913	150,991
Bartlett's Test p–value	0.000	0.000	0.000

In this study, the number of retained factors was determined based on variance-explained thresholds (Table 4): 95% for Cluster 1 and Cluster 3, and 90% for Cluster 2. Accordingly, 28, 24, and 27 factors were retained for Clusters 1, 2, and 3, respectively. Taking Cluster 2 as an example, Table 3 presents the variance explained by successive factor extraction. The first 12 factors exhibit eigenvalues greater than 1 and collectively account for approximately 62.55% of the total variance. When all 24 factors are included, the cumulative explained variance reaches 90.67%, indicating that the extracted factors effectively capture the majority of the information in the original variables.

Table 4. Variance Explained by Extracted Factors in Cluster 2.

Factor	Eigenvalue	Proportional Variance	Cumulative Variance
Factor_1	4.9028	0.1362	0.1362
Factor_2	3.6668	0.1019	0.2380
Factor_3	2.1857	0.0607	0.2988
Factor_4	1.9370	0.0538	0.3526
Factor_5	1.8217	0.0506	0.4032
Factor_6	1.3719	0.0381	0.4413
Factor_7	1.2226	0.0340	0.4752
Factor_8	1.1849	0.0329	0.5081
Factor_9	1.1210	0.0311	0.5393

 doi:10.20944/preprints202603.0178.v1

Factor_10	1.0451	0.0290	0.5683
Factor_11	1.0356	0.0288	0.5971
Factor_12	1.0247	0.0285	0.6255
Factor_13	0.9992	0.0278	0.6533
Factor_14	0.9920	0.0276	0.6809
Factor_15	0.9693	0.0269	0.7078
Factor_16	0.9350	0.0260	0.7338
Factor_17	0.9298	0.0258	0.7596
Factor_18	0.8981	0.0249	0.7845
Factor_19	0.8796	0.0244	0.8090
Factor_20	0.8379	0.0233	0.8322
Factor_21	0.7538	0.0209	0.8532
Factor_22	0.7000	0.0194	0.8726
Factor_23	0.6320	0.0176	0.8902
Factor_24	0.5935	0.0165	0.9067

The heatmap (Figure 5) visualizes factor loadings after oblique (Promax) rotation, where each row represents an original variable and each column corresponds to one of the 27 retained factors. Color intensity and direction (blue: negative, red: positive) indicate the strength and sign of the association between variables and factors. Numbers overlaid on the cells denote the exact loading values.

This representation illustrates how the underlying latent constructs—such as locational attributes (e.g., Business_District, Administrative_Region), structural features (e.g., Internal_Area, Building_Area), and market—dynamic indicators (e.g., Listing_Price, Page_Views)—are captured by the rotated factor solution. A small number of loadings exceed |1| (e.g., Business_District reaches −0.62), reflecting the slight Heywood—case occurrence discussed in the text, yet the overall pattern supports the interpretability and predictive utility of the extracted factor structure.

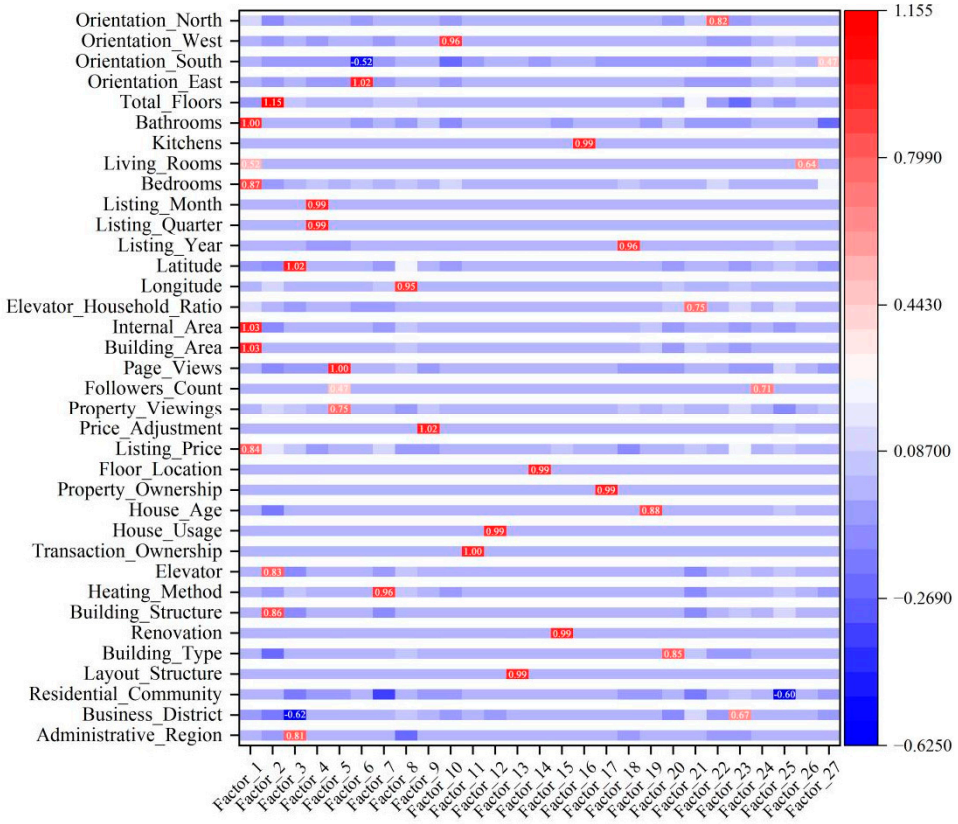


Figure 5. Rotated factor pattern matrix for the 27 extracted factors in Cluster 2.

 doi:10.20944/preprints202603.0178.v1

After determining the number of factors, oblique rotation was applied to the extracted factors. Table 3 presents the partial rotated factor pattern matrix for Cluster 2 following oblique rotation.

During factor analysis, it was observed that the estimated communality of some factors slightly exceeded 1 (i.e., a "Heywood case"). This typically occurs when the model attempts to capture maximum information from the variables or when the sample size is relatively small compared to the number of variables. Although such estimates are not optimal from a strict measurement-model perspective, the following considerations supported the retention of this factor structure:

First, the magnitude of the deviation was minimal (maximum absolute deviation < 0.2). Second, and more importantly, the factor scores generated from this factor structure demonstrated good initial predictive validity even in a single model (e.g., Random Forest achieved $R^2 = 0.877$ in Cluster 1 and $R^2 = 0.883$ in Cluster 2). These results indicate that, despite being a statistical boundary case, the extracted factors possess practical utility for prediction.

Therefore, this factor structure was retained for subsequent ensemble modeling, with the primary objective being the maximization of predictive accuracy rather than the validation of a perfect Measurement model.

3.2.5. Lasso Feature Selection

To further refine the feature set from the extracted composite factors for housing price prediction—aiming to avoid overfitting and enhance model interpretability—this study employed Lasso regression for feature selection[45]. The Lasso method imposes an L1 regularization penalty on regression coefficients, automatically shrinking the coefficients of less important variables to zero, thereby performing embedded feature selection.

Using the LassoCVfunction with 10-fold cross-validation, the optimal regularization strength was determined (e.g., for Cluster 2, $\alpha = 0.432$). Figure 6 displays the absolute standardized coefficients of each factor under the optimal Lasso model, where the magnitude reflects the relative importance of the factor to the prediction target. As shown, the importance varies significantly among factors: Factor_1 and Factor_5 contribute the most, whereas some factors show a negative association. This selection process ensures that only features with significant predictive power are retained for the final predictive model.

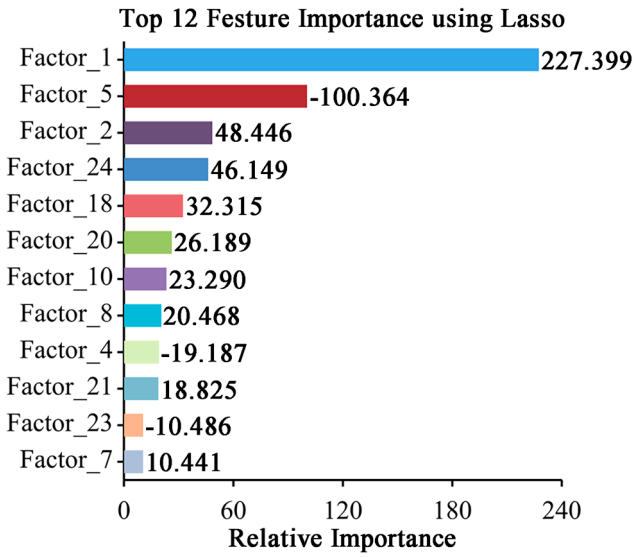


Figure 6. Feature Importance Bar Chart.

Figure 7 plots the coefficient paths of each factor as the regularization strength varies, visually illustrating how features are progressively eliminated and verifying the stability of the selection process.

 doi:10.20944/preprints202603.0178.v1

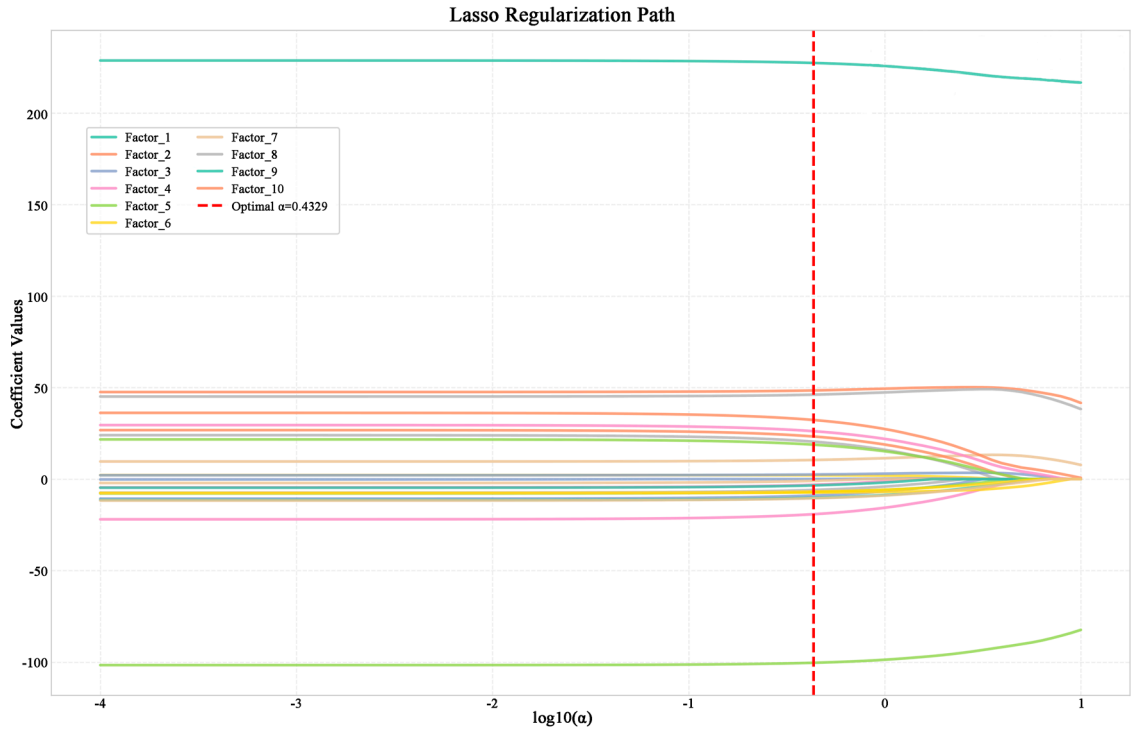


Figure 7. Regularization Path Map.

Ultimately, all factors with non-zero coefficients at the optimal α were retained, forming a concise yet highly predictive feature set for the final housing price prediction model. After applying Lasso feature selection, 24, 24, and 27 factors were retained for Clusters 1, 2, and 3, respectively. The results indicate that in Clusters 2 and 3, all post-factor-analysis factors possess significant predictive power, whereas Cluster 1 contains a small number of less predictive factors, underscoring the importance of feature selection in certain scenarios.

3.3. Model Performance Comparison

3.3.1. Factor Naming

To facilitate the interpretability of the factor variables, this study labeled and interpreted the first five extracted factors. The resulting factor definitions are summarized in Table 5.

- Scale Value Factor: Defined by the listing price, area, and the number of rooms, it reflects the value foundation formed by the physical size of the property
- Architectural Feature Factor: Defined by the building type, structure, elevator availability, and floor level, it captures the hardware attributes and construction quality of the property.
- Scale—Driven Attention Factor: Defined by the viewing frequency, the number of interested buyers, and the area, it reveals how the property's basic scale drives market attention.
- Listing—Time Periodicity Factor: Defined by the listing quarter and month, it captures seasonal patterns in market activity.
- Core—Ideal—Location Factor: Defined by the business—district grade and longitude, it identifies advantageous locations characterized by commercial—resource concentration and spatial scarcity.

These factors are particularly meaningful for Cluster 2, which represents Beijing's core urban districts (Dongcheng, Xicheng, Haidian, and Chaoyang). In this high—value sub—market, fundamental attributes such as scale and architectural features become critical valuation dimensions. The high market attention and concentrated improvement—oriented demand make the driving effect of the area on attention more pronounced. The mature market is also susceptible to seasonal supply—and—demand fluctuations. Moreover, the pronounced spatial differentiation in location value means

 doi:10.20944/preprints202603.0178.v1

that commercial resources and spatial scarcity jointly determine the externality—based premium. Therefore, this factor system comprehensively covers the main dimensions of housing—price formation in core urban districts, and the analytical results are well—grounded in reality.

Table 5. Naming and Interpretation of the First Five Factors.

Factor	Factor Name
Factor 1	Scale Value Factor
Factor 2	Architectural Feature Factor
Factor 3	Scale-Driven Attention Factor
Factor 4	Listing-Time Periodicity Factor
Factor 5	Core-Ideal-Location Factor

3.3.2. Model Performance Comparison

To comprehensively evaluate the predictive performance across different market segments, a Stacking ensemble framework was implemented using Linear Regression (LR) and Decision Tree (DT) as meta-models. All base models were fine-tuned via Bayesian optimization, and meta-features were generated using 5-fold cross-validation. Table 6 summarizes the final model performance metrics for the three clusters, while Figure 8 (Radar-Style Performance Diagram) provides a visual comparison of model performance across Cluster 1 (suburban market), Cluster 2 (central-urban market), and Cluster 3 (sub-central market).

Table 6. Performance Comparison of Stacking Models Across Clusters.

Cluster	Meta-Model	RMSE	MAE	R^2	MAPE (%)	Improvement Over Best Base Model*
Cluster 1	LinearRegression	49.20	35.32	0.919	9.00	−0.04%
	DecisionTree	51.03	36.27	0.913	9.24	−0.71%
Cluster 2	LinearRegression	99.17	68.61	0.916	10.69	+0.56%
	DecisionTree	103.36	70.00	0.909	10.93	−0.23%
Cluster 3	LinearRegression	33.51	24.25	0.902	9.15	+1.29%
	DecisionTree	36.52	26.59	0.883	10.08	−0.77%

Note: Improvement is calculated relative to the best-performing single base model in each cluster (LightGBM for Cluster 1 and Cluster 3; XGBoost for Cluster 2).

Figure 8 illustrates the relative performance of different models across three market clusters in terms of R^2 and RMSE. In all clusters, the Stacking—LR and Stacking—DT models consistently outperform or are competitive with the best individual base models (ElasticNet, LightGBM, RandomForest, and XGBoost). Specifically:

(1) Cluster 1 (High—Volume Suburban Market; N = 78,417)

The Linear Regression (LR) stacking model achieved an R^2 of 0.9193, performing virtually identically to the best single base model, LightGBM (R^2 = 0.9197; –0.04% relative change). This result is encouraging: when the data volume is large and the feature–target relationship can be well captured by a single complex model, the stacking framework at least maintains the same predictive accuracy. Moreover, by blending linear (ElasticNet) and nonlinear (tree—based) base—model outputs, the stacked predictions are likely to be more robust and generalizable, even if the out—of—sample metrics did not show a clear uplift in the present test set.

(2) Cluster 2 (Central—Urban Market; Log—Transformed)

In this sub—market, the LR stacking model (R^2 = 0.9162) outperformed all individual base models, delivering a modest but consistent improvement of +0.56% over the best single model, XGBoost. The positive gain suggests that the ensemble captured complementary predictive signals that were missed by any single learner. Notably, the target variable (housing price) in Cluster 1 exhibited pronounced positive skewness (1.059), and a log transformation was applied to satisfy

 doi:10.20944/preprints202603.0178.v1

distributional assumptions. Here, stacking effectively integrated the diverse predictive perspectives of XGBoost, LightGBM, and RandomForest, yielding higher accuracy.

(3) Cluster 3 (Small—Sample Sub—Central Market; N = 9,507)

The ensemble advantage was most pronounced in this cluster. Despite the limited sample size and relatively high price variability, the LR stacking model achieved the largest relative improvement (+1.29%), raising R^2 from 0.8903 (LightGBM) to 0.9018. This outcome strongly indicates that in settings characterized by long—term temporal spans, limited data, noticeable noise, or elevated overfitting risk, stacking multiple base models helps balance bias and variance, thereby enhancing generalization—a regularization effect clearly demonstrated here.

Across all three clusters, stacking with a simple linear—regression meta—model consistently outperformed stacking with a decision—tree meta—model. This supports a key methodological insight: when the base models are already powerful nonlinear learners (e.g., gradient—boosted trees, random forests), the residuals or synergies among their predictions tend to exhibit an approximately linear structure. Using a complex nonlinear meta—model (e.g., decision tree) for secondary learning may not capture additional meaningful signals and can instead introduce unnecessary variance, leading to degraded performance. Hence, the configuration "strong nonlinear base models + simple linear meta—model" proved to be the most accurate and robust stacking architecture in this study—a finding that aligns with the principle of Occam's Razor in machine learning.

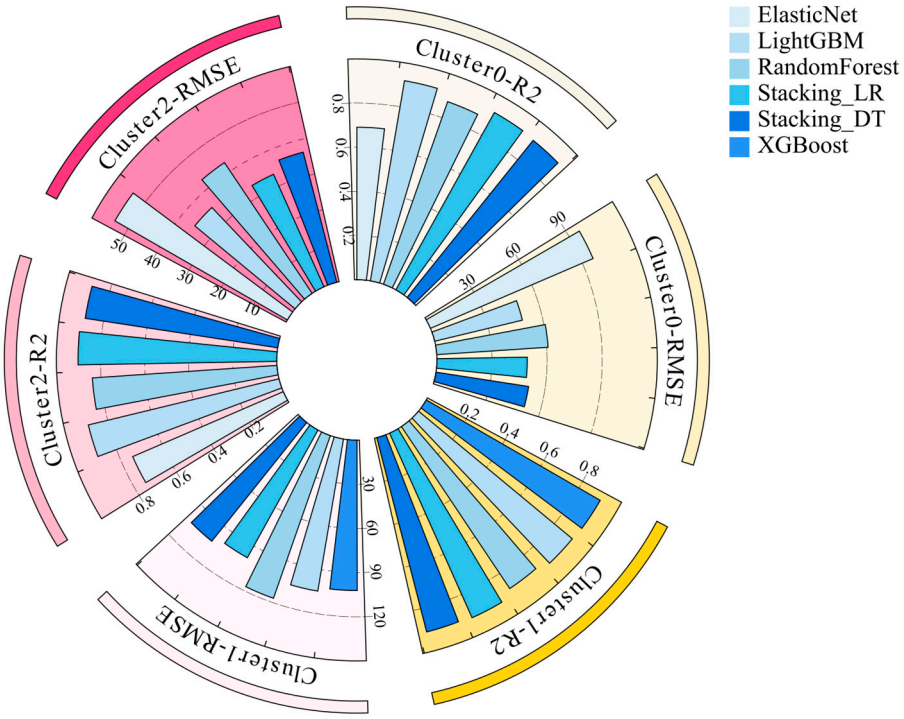


Figure 8. Stacking ensemble model performance across submarkets.

3.4. Feature Importance Analysis

To thoroughly understand the contribution and economic implications of each composite factor in predicting housing prices, we conducted an interpretability analysis of the optimal Linear Regression (LR) Stacking model using SHAP (SHapley Additive exPlanations) values. The results were further integrated with findings from the preceding factor analysis.

Figure 9 presents the SHAP feature importance and directional impacts of the LR Stacking model for Cluster 2.

To delve deeper into the mechanisms through which the composite factors influence price prediction, we compared the results of Lasso—based feature selection with SHAP analysis applied to the optimal stacking model. In Cluster 2 (central urban area), both methods revealed generally

 doi:10.20944/preprints202603.0178.v1

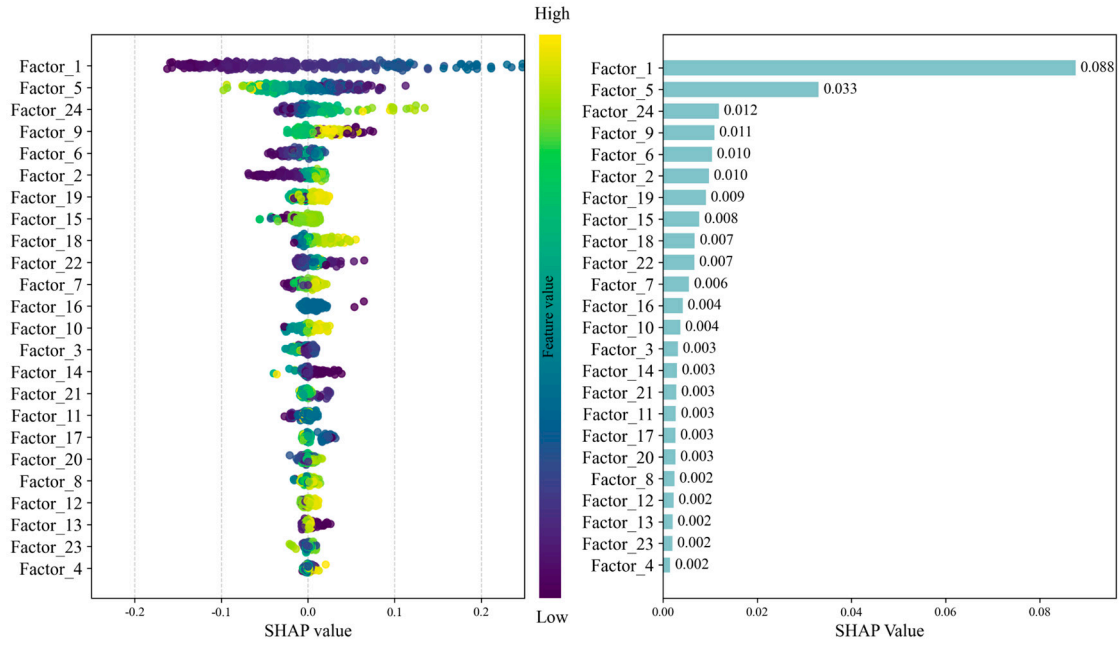
consistent yet nuanced differences in factor importance rankings, offering complementary perspectives for interpreting the underlying drivers.

A comparative analysis was conducted based on the ranking of key factors according to the absolute Lasso coefficients and the mean absolute SHAP values. Both methods identified the "Scale Value Factor" (Factor 1) and the "Core—Ideal—Location Factor" (Factor 5) as the two most crucial predictive dimensions. However, their relative order of importance varied: Lasso regression ranked the Scale Value Factor first, while SHAP analysis ascribed the highest contribution to the Core—Ideal—Location Factor.

This discrepancy has methodological implications. Lasso, being a linear model, estimates coefficients that directly measure the unique contribution of each variable within a linear predictive framework. Its prioritization of the Scale Value Factor may suggest a stable and strong linear association between this factor and housing prices, making it an essential baseline predictor in linear modeling. Nevertheless, it is worth noting that in Cluster 2, purely linear models like Lasso and Elastic Net showed very low predictive accuracy and were ultimately replaced by the XGBoost base model in the stacking ensemble. Therefore, the feature importance derived from Lasso coefficient magnitudes should be interpreted with caution and mainly serves as a reference, highlighting the necessity of complementary SHAP—based analysis.

In contrast, SHAP values are derived from the complex, non—linear ensemble (Stacking) model and can capture interactive and non—linear effects. The elevation of the Core—Ideal—Location Factor to the top position in the SHAP ranking indicates that this factor has a significant impact on housing prices through a more intricate—and arguably more realistic—mechanism, which may be underestimated in simple additive linear models. Its prominence in the SHAP analysis is well—aligned with the established economic understanding that location is a crucial determinant of property value, especially in a high—stakes, spatially differentiated market such as Beijing's central urban areas.

In summary, the integration of traditional statistical dimensionality reduction (factor analysis) with modern interpretable machine learning techniques (Lasso and SHAP) allows for a multi—layered interpretation—from "variable structure" to "predictive utility." This approach not only validates the predictive relevance of core economic constructs (scale, location, and architectural attributes) but also, through cross—method comparison and triangulation, deepens our understanding of how different factors influence housing prices via either linear or non—linear pathways. Consequently, it enhances the robustness and insightful depth of the model's conclusions.



 doi:10.20944/preprints202603.0178.v1

Figure 9. presents the SHAP feature importance and directional impacts of the LR Stacking model for Cluster 2.

4. Conclusions and Future Directions

This study addresses the high spatial heterogeneity and nonlinear pricing mechanisms in urban second—hand housing markets by proposing a predictive and interpretative framework that integrates spatial clustering and partitioned stacking ensemble learning. This approach advances real estate valuation modeling in three key aspects:

The study effectively combines unsupervised spatial clustering with supervised stacked ensemble modeling, implementing a "partition—first, model—second" framework design. This design not only identifies and adapts to the price formation mechanisms of different sub—markets but also overcomes the limitations of traditional GWR models in capturing complex nonlinear relationships. The results demonstrate that across all sub—markets, the stacked ensemble architecture with linear regression as the meta—model generally outperforms single machine learning models and ensemble strategies with decision trees as meta—models, confirming the robustness and effectiveness of "strong nonlinear base models + simple linear meta—model" in complex prediction tasks.

By extracting economically meaningful common factors from high—dimensional features through factor analysis, combining Lasso regression for feature screening, and deconstructing predictive contributions using SHAP values, the framework achieves a progressive analysis from "feature dimensionality reduction" to "factor interpretation" and finally to "mechanism revelation." The study finds that in the core urban sub—market, the "core ideal location factor" plays a dominant role through nonlinear mechanisms, while the "scale value factor" exhibits a stable linear association. This finding deepens the understanding of the pathways through which different factors operate and provides a basis for differentiated policy design.

The proposed framework can be directly applied to scenarios such as automated real estate valuation systems, regional market monitoring, and policy effect simulation. Through empirical analysis of Beijing, the effectiveness of the framework in identifying sub—market structures, capturing nonlinear characteristics of locational premiums, and improving predictive accuracy has been demonstrated. The framework exhibits strong scalability and can be adapted to different urban structures, data foundations, and business needs, providing decision—support tools with both predictive performance and interpretative depth for financial institutions, government departments, and market participants.

Future research could explore the following directions: first, incorporating temporal dimensions to construct dynamic spatial panel models capturing spatiotemporal interaction effects in price evolution; second, integrating multi—modal data such as text and images to further enrich feature representation; and third, conducting external validation in more cities with diverse market structures to test the generalizability and robustness of the framework.

Author Contributions: conceptualization, Zezhong Wang; investigation, Jing Li, Wanxin Li,and Zezhong Wang; methodology, Wanxin Li and Shuohan Jiang; writing—original draft preparation, Jing Li; writing—review and editing, Xiaolin Sun; project administration,Jing Li; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

 doi:10.20944/preprints202603.0178.v1

References

1. Zali, S.; Pahlavani, P.; Ghorbanzadeh, O.; Khazravi, A.; Ahmadlou, M.; Givekesh, S. Housing Price Modeling Using a New Geographically, Temporally, and Characteristically Weighted Generalized Regression Neural Network (GTCW-GRNN) Algorithm. *BUILDINGS-BASEL* **2025**, *15*, doi:10.3390/buildings15091405.
2. Weinstock, L.R. Introduction to U.S. Economy: Housing Market.
3. Ren, J.; Gao, X. Grid Density Algorithm-Based Second-Hand Housing Transaction Activity and Spatio-Temporal Characterization: The Case of Shenyang City, China. *Isprs Int. J. Geo-inf.* **2024**, *13*, doi:10.3390/ijgi13080286.
4. Bitter, C.; Mulligan, G.F.; Dall'erba, S. Incorporating Spatial Variation in Housing Attribute Prices: A Comparison of Geographically Weighted Regression and the Spatial Expansion Method. *J. Geogr. Syst.* **2007**, *9*, 7–27, doi:10.1007/s10109-006-0028-7.
5. Wheeler, D.; Tiefelsdorf, M. Multicollinearity and Correlation among Local Regression Coefficients in Geographically Weighted Regression. *J. Geogr. Syst.* **2005**, *7*, 161–187, doi:10.1007/s10109-005-0155-6.
6. Zhao, C.; Liu, F. Impact of Housing Policies on the Real Estate Market—Systematic Literature Review. *Heliyon* **2023**, *9*, e20704, doi:10.1016/j.heliyon.2023.e20704.
7. Ding, J.; Cen, W.; Wu, S.; Chen, Y.; Qi, J.; Huang, B.; Du, Z. A Neural Network Model to Optimize the Measure of Spatial Proximity in Geographically Weighted Regression Approach: A Case Study on House Price in Wuhan. *Int. J. Geogr. Inf. Sci.* **2024**, *38*, 1315–1335, doi:10.1080/13658816.2024.2343771.
8. Yin, Z.; Sun, R.; Bi, Y. Spatial-Temporal Change Trend Analysis of Second-Hand House Price in Hefei Based on Spatial Network. *Comput. Intell. Neurosci.* **2022**, *2022*, doi:10.1155/2022/6848038.
9. Wu, G.; Guo, W.; Niu, X. Spillover Effect Analysis of Home-Purchase Limit Policy on Housing Prices in Large and Medium-Sized Cities: Evidence from China. *PLoS One* **2023**, *18*, doi:10.1371/journal.pone.0280235.
10. Tekouabou, S.C.K.; Gherghina, S.C.; Kameni, E.D.; Filali, Y.; Idrissi Gartoumi, K. AI-Based on Machine Learning Methods for Urban Real Estate Prediction: A Systematic Survey. *Arch. Comput. Methods Eng.* **2024**, *31*, 1079–1095, doi:10.1007/s11831-023-10010-5.
11. Zhang, J.; Liu, Z. Interval Prediction of Crude Oil Spot Price Volatility: An Improved Hybrid Model Integrating Decomposition Strategy, IESN and ARIMA. *Expert Systems with Applications* **2024**, *252*, 124195, doi:10.1016/j.eswa.2024.124195.
12. Pei, M.; Gong, R.; Ye, L.; Chen, L.; Sun, Y.; Tang, Y. Spatiotemporal Sparse Autoregressive Distributed Lag Model with Extended Regressors for Regional Wind Power Forecasting. *Appl. Energy* **2026**, *404*, 127205, doi:10.1016/j.apenergy.2025.127205.
13. Kumari, P.; Goswami, V.; Harshith, N.; Pundir, R.S. Recurrent Neural Network Architecture for Forecasting Banana Prices in Gujarat, India. *PLoS One* **2023**, *18*, doi:10.1371/journal.pone.0275702.
14. Chen, C.W.S.; Chiu, L.M. Ordinal Time Series Forecasting of the Air Quality Index. *Entropy* **2021**, *23*, doi:10.3390/e23091167.
15. Zhang, C.; Chen, K. Unravelling the Interplay of Crude Oil, Renewable Energy, and Commodity Price Volatility: A DCC-GARCH Model Approach on the Chinese Stock Market. *Renewable Energy* **2026**, *256*, 124128, doi:10.1016/j.renene.2025.124128.
16. Li, Z.; Xie, S.; Zhang, Y.; Hu, J. A Study on House Price Prediction Based on Stacking-Sorted-Weighted-Ensemble Model. *J. Internet Technol.* **2022**, *23*, 1139–1146, doi:10.53106/160792642022092305022.
17. Mao, Y.; Duan, Y.; Guo, Y.; Wang, X.; Gao, S.; Ali, G. A Study on the Prediction of House Price Index in First-Tier Cities in China Based on Heterogeneous Integrated Learning Model. *J. Math.* **2022**, *2022*, doi:10.1155/2022/2068353.
18. Rey-Blanco, D.; Zofío, J.L.; González-Arias, J. Improving Hedonic Housing Price Models by Integrating Optimal Accessibility Indices into Regression and Random Forest Analyses. *Expert Syst. Appl.* **2024**, *235*, 121059, doi:10.1016/j.eswa.2023.121059.
19. Amjad, M.; Ahmad, I.; Ahmad, M.; Wroblewski, P.; Kaminski, P.; Amjad, U. Prediction of Pile Bearing Capacity Using XGBoost Algorithm: Modeling and Performance Evaluation. *Appl. Sci.-basel* **2022**, *12*, doi:10.3390/app12042126.

20. Dong, J.; Chen, Y.; Yao, B.; Zhang, X.; Zeng, N. A Neural Network Boosting Regression Model Based on XGBoost. *Appl. Soft Comput.* **2022**, *125*, doi:10.1016/j.asoc.2022.109067.
21. Simarmata, N.; Wikantika, K.; Tarigan, T.A.; Aldyansyah, M.; Tohir, R.K.; Fauzi, A.I.; Fauzia, A.R. Comparison of Random Forest, Gradient Tree Boosting, and Classification and Regression Trees for Mangrove Cover Change Monitoring Using Landsat Imagery. *The Egyptian Journal of Remote Sensing and Space Sciences* **2025**, *28*, 138–150, doi:10.1016/j.ejrs.2025.02.002.
22. Wang, J.; Ji, H.; Wang, L. Forecasting Second-Hand House Prices in China Using the GA-PSO-BP Neural Network Model. *PLoS One* **2025**, *20*, doi:10.1371/journal.pone.0322821.
23. Rampini, L.; Re Cecconi, F. Artificial Intelligence Algorithms to Predict Italian Real Estate Market Prices. *JPIF* **2022**, *40*, 588–611, doi:10.1108/JPIF-08-2021-0073.
24. Yu, B.; Yan, D.; Wu, H.; Wang, J.; Chen, S. A Novel Prediction Model for the Sales Cycle of Second-Hand Houses Based on the Hybrid Kernel Extreme Learning Machine Optimized Using the Improved Crested Porcupine Optimizer. *Buildings* **2025**, *15*, doi:10.3390/buildings15071200.
25. Shao, J.; Yu, L.; Zeng, N.; Hong, J.; Wang, X. A Multi-Scale Analysis Method with Multi-Feature Selection for House Prices Forecasting. *Applied Soft Computing* **2025**, *171*, 112779, doi:10.1016/j.asoc.2025.112779.
26. Huang, M.; Liu, D.; Ma, L.; Wang, J.; Wang, Y.; Chen, Y. A Prediction Method of Electromagnetic Environment Effects for UAV LiDAR Detection System. *Complexity* **2021**, *2021*, doi:10.1155/2021/7190446.
27. Luo, Z.; Wang, H.; Li, S. Prediction of International Roughness Index Based on Stacking Fusion Model. *Sustainability* **2022**, *14*, doi:10.3390/su14126949.
28. Guo, Y.; Liu, H.; Zhou, X.; Chen, J.; Guo, L. Research on Coal and Gas Outburst Risk Warning Based on Multiple Algorithm Fusion. *Appl. Sci.-basel* **2023**, *13*, doi:10.3390/app132212283.
29. Chen, H.; Zhang, X. Path Planning for Intelligent Vehicle Collision Avoidance of Dynamic Pedestrian Using Att-LSTM, MSFM, and MPC at Unsignalized Crosswalk. *IEEE Trans. Ind. Electron.* **2022**, *69*, 4285–4295, doi:10.1109/TIE.2021.3073301.
30. Chen, H.; Zhang, X.; Yang, W.; Lin, Y. A Data-Driven Stacking Fusion Approach for Pedestrian Trajectory Prediction. *Transportmetrica B-Transp. Dyn.* **2023**, *11*, 548–571, doi:10.1080/21680566.2022.2103050.
31. Matsunaga, M. How to Factor-Analyze Your Data Right: Do's, Don'ts, and How-to's. *International Journal of Psychological Research* **2010**, *3*, 97–110, doi:10.21500/20112084.854.
32. Williams, J.S. Review of the Essentials of Factor Analysis. *Contemporary Sociology* **1974**, *3*, 411–411, doi:10.2307/2061984.
33. Hu, L.; Gao, L.; Li, Y.; Zhang, P.; Gao, W. Feature-Specific Mutual Information Variation for Multi-Label Feature Selection. *Inf. Sci.* **2022**, *593*, 449–471, doi:10.1016/j.ins.2022.02.024.
34. Enwere, K.; Nduka, E.; Ogoke, U. Comparative Analysis of Ridge, Bridge and Lasso Regression Models in the Presence of Multicollinearity. *IPS Intelligentsia Multidisciplinary Journal* **2023**, *3*, 1–8, doi:10.54117/iimj.v3i1.5.
35. Fryer, D.; Strümke, I.; Nguyen, H. Shapley Values for Feature Selection: The Good, the Bad, and the Axioms Available online: https://arxiv.org/abs/2102.10936v1 (accessed on 15 February 2026).
36. Zhang, X.; Dai, C.; Li, W.; Chen, Y. Prediction of Compressive Strength of Recycled Aggregate Concrete Using Machine Learning and Bayesian Optimization Methods. *Front. Earth Sci.* **2023**, *11*, doi:10.3389/feart.2023.1112105.
37. Tao, S.; Peng, P.; Li, Y.; Sun, H.; Li, Q.; Wang, H. Supervised Contrastive Representation Learning with Tree-Structured Parzen Estimator Bayesian Optimization for Imbalanced Tabular Data. *Expert Syst. Appl.* **2024**, *237*, 121294, doi:10.1016/j.eswa.2023.121294.
38. Wang, H.; Liang, Q.; Hancock, J.T.; Khoshgoftaar, T.M. Feature Selection Strategies: A Comparative Analysis of SHAP-Value and Importance-Based Methods. *J Big Data* **2024**, *11*, 44, doi:10.1186/s40537-024-00905-w.
39. Aldrees, A.; Khan, M.; Taha, A.T.B.; Ali, M. Evaluation of Water Quality Indexes with Novel Machine Learning and SHapley Additive ExPlanation (SHAP) Approaches. *J. Water Process Eng.* **2024**, *58*, 104789, doi:10.1016/j.jwpe.2024.104789.
40. Huang, P.; Cai, J.; Wang, J.; Chen, H.; Zhang, P. High-Accuracy ETA Prediction for Long-Distance Tramp Shipping: A Stacked Ensemble Approach. *J. Mar. Sci. Eng.* **2026**, *14*, doi:10.3390/jmse14020177.

 doi:10.20944/preprints202603.0178.v1

41. Alsuqayh, N.; Mirza, A.; Alhogail, A. Exploring Feature Engineering and Explainable AI for Phishing Website Detection: A Systematic Literature Review. *International Journal of Electrical and Computer Engineering (IJECE)* **2025**, *15*, 5863–5878, doi:10.11591/ijece.v15i6.pp5863-5878.
42. B, H.; Mb, H.; M, R. Psychometric Evaluation of the Perfectionism Scale's Characteristics Regarding Physical Appearance in Patients Seeking Rhinoplasty Surgery. *JPRAS open* **2024**, *41*, doi:10.1016/j.jpra.2024.05.004.
43. Al-Hadeedy, I.Y.; Ameen, Q.A.; Shaker, A.S.; Mohamed, A.H.; Taha, M.W.; Hussein, S.M. Using the Principal Component Analysis of Body Weight in Three Genetic Groups of Japanese Quail. *Iop Conf. Ser.: Earth Environ. Sci* **2023**, *1252*, 012148, doi:10.1088/1755-1315/1252/1/012148.
44. Užarević, Z.; Petković, F.; Koruga, A.S.; Kampić, I.; Popijač, Ž.; Soldo, S.B. The Multiple Sclerosis Intimacy and Sexuality Questionnaire-15: Validity, Reliability, and Factor Structure of Croatian Version. *Journal of Health Sciences* **2025**, *15*, 114–118, doi:10.17532/jhsci.2025.2877.
45. Hu, J.-Y.; Wang, Y.; Tong, X.-M.; Yang, T. When to Consider Logistic LASSO Regression in Multivariate Analysis? *Eur J Surg Oncol* **2021**, *47*, 2206, doi:10.1016/j.ejso.2021.04.011.