International Journal of Engineering Research & Science (IJOER)
ISSN: [2395-6992]
[Vol-12, Issue-8, August - 2026]
Website: www.ijoer.com
Journal DOI: 10.25125/engineering-journal

House Price Prediction in Nanjing Based on Machine Learning and SHAP Interpretability

Yang Zaibin[1]; Cui Xinxin[2*]

[1,2]School of Mathematics and Statistics, Guizhou University of Finance and Economics, Guiyang 550025, China.
[1,2]School of Big Data Statistics, Guizhou University of Finance and Economics, Guiyang 550025, China.
*Corresponding Author

Received: 24 July 2026/ Revised: 01 August 2026/ Accepted: 11 August 2026/ Published: 15-08-2026

Abstract— *Accurate house price prediction is of great significance for homebuyers, developers, and policymakers, yet the complexity and non-linearity of housing markets pose substantial challenges to traditional linear models. This study addresses the dual goals of high predictive accuracy and high interpretability by systematically comparing five regression models—linear regression, random forest, XGBoost, LightGBM, and multilayer perceptron (MLP)—on a large-scale second-hand housing dataset from Nanjing, China. Model performance is evaluated using MAE, RMSE, and R^2. Results show that ensemble tree models significantly outperform linear regression and MLP, with LightGBM and XGBoost achieving the lowest MAE (6,200 RMB/m²) and RMSE (8,800 RMB/m²), and LightGBM attaining the highest R^2 (0.92). We further employ SHAP (SHapley Additive exPlanations) to open the "black box" of the optimal LightGBM model. Global feature importance reveals that house age, subway distance, and decoration condition are the three most influential drivers of Nanjing house prices. SHAP summary plots demonstrate that newer houses, closer subway proximity, and better decoration consistently raise prices, while area exhibits a non-monotonic effect—small units command higher unit prices under total-price constraints, whereas very large luxury properties show diminishing marginal value. Prediction diagnostics confirm robust performance in the mainstream price range (15,000–35,000 RMB/m²), though extreme high-end properties remain harder to predict due to unobserved idiosyncratic factors. This study not only provides a high-performance predictive tool for the Nanjing market but also offers transparent, actionable insights into the underlying price-formation mechanism. The integrated framework of machine learning plus SHAP interpretability is readily generalisable to other cities and real-estate contexts, supporting evidence-based decision-making and policy design.*

Keywords— House price prediction; Machine learning;

I. INTRODUCTION

Housing issues are central to national welfare and people's livelihoods, and the stable operation of the real estate market plays a pivotal role in economic and social development. As a major central city in the Yangtze River Delta, Nanjing has witnessed persistently active real estate markets in recent years, with house prices influenced by multiple intertwined factors, exhibiting pronounced regional differentiation and volatility. Against this backdrop, constructing a scientific and accurate house price prediction model not only provides a basis for homebuyer decision making and developer pricing, but also assists government agencies in grasping market dynamics and refining regulatory measures. This holds considerable practical significance and research value.

Research on house price prediction can be traced back to the hedonic price theory proposed by Rosen [16], which posits that the value of a heterogeneous good can be explained by the implicit prices of its constituent attributes. This framework became the cornerstone of a large body of empirical work. Building on this foundation, traditional approaches predominantly employ multiple linear regression to estimate the marginal effects of individual features, offering strong interpretability. However, real world house price formation often involves complex nonlinear interactions—for example, the synergy between floor area and

location, or the non-monotonic influence of floor level. Linear models inherently struggle to capture such relationships, and their predictive accuracy often falls short of growing practical demands [3].

With advances in computational power and the proliferation of data sources, machine learning models have gradually been introduced into real estate valuation. Selim [18] compared the predictive performance of artificial neural networks and hedonic price models on Turkish housing data, finding that neural networks yielded significantly lower prediction errors than traditional regression, thereby pioneering the application of nonlinear models in this domain. Subsequently, various ensemble learning and gradient boosting methods have been widely explored for house price prediction: Breiman [1] proposed random forests, which alleviate overfitting through bootstrap aggregation and random feature selection; Friedman [7] established the general framework of gradient boosting decision trees; XGBoost by Chen and Guestrin [2] and LightGBM by Ke et al. [10] further optimised the efficiency and accuracy of gradient boosting, demonstrating exceptional performance in structured data regression tasks. Meanwhile, simple deep learning networks such as multilayer perceptrons have offered another nonlinear benchmark [4]. Nevertheless, these high accuracy models are often regarded as "black boxes" whose internal decision making processes are not intuitively understandable, which to some extent constrains their trustworthiness and acceptance in high stakes decisions [5]. Recent systematic reviews have consolidated these advances, revealing that Random Forest, Gradient Boosting Machine, XGBoost, and LightGBM are the most frequently employed models in real estate price prediction, consistently outperforming traditional linear approaches. However, the proliferation of these high-accuracy models has intensified a fundamental tension: while ensemble methods excel at capturing complex nonlinearities and interaction effects, their internal decision logic remains opaque, prompting scholars to call for integrating explainable artificial intelligence techniques into mass real estate valuation systems [6].

In response, researchers have increasingly turned to SHAP to bridge the gap between predictive accuracy and interpretability. For instance, Hu et al. [8] combined XGBoost with SHAP to examine the nonlinear effects of public service amenities and street-view greenery on housing prices in Shanghai, revealing that residents paid a premium of approximately 2,000 yuan/m² for higher green views [9]. Wang and Li [19] employed Random Forest with SHAP to quantify feature contributions in a rapidly urbanizing Chinese city, identifying location-based factors and environmental attributes as the most influential drivers. Ye [20] demonstrated that XAI approaches should be systematically integrated into mass real estate appraisal and urban housing research more broadly. Internationally, studies on Seoul apartment transactions and Belgian residential rents have similarly validated the effectiveness of SHAP for extracting transparent, actionable insights from tree-based models.

The advantages of SHAP over alternative interpretability methods are substantial. Unlike LIME, which provides only local approximations around individual predictions, SHAP offers both global and local explanations with desirable theoretical properties—local accuracy, consistency, and missingness—making it particularly suitable for high-stakes real estate decision-making [10]. The TreeSHAP algorithm further enables efficient and exact computation for tree-based ensembles, which is precisely the algorithmic family that dominates structured data regression tasks [11]. Moreover, recent research has extended SHAP to capture spatial heterogeneity and nonlinear threshold effects, demonstrating its versatility across diverse urban contexts.

Ribeiro et al. [15] proposed LIME (Local Interpretable Model-agnostic Explanations), which approximates the decision boundary of a complex model using a local surrogate model. Building on this foundation and seeking a more unified and theoretically grounded framework, Lundberg and Lee [11] introduced SHAP based on Shapley values from cooperative game theory, assigning each feature a unified contribution to the prediction, with desirable properties including local accuracy, consistency, and missingness. In particular, the TreeSHAP algorithm for tree-based ensembles enables efficient and exact computation of SHAP values, providing a powerful tool for opening the "black box" of complex models. Although SHAP has been successfully applied in fields such as healthcare and credit scoring, its use in house price prediction—especially for the Nanjing market—remains relatively scarce.

Existing domestic house price prediction literature mostly focuses on first tier cities such as Beijing and Shanghai, and most studies stop at comparing model prediction accuracy, with limited in depth interpretation of how features affect predictions [13]. As a key central city in eastern China, the driving factors and mechanisms of house prices in Nanjing warrant dedicated investigation [14]. Therefore, this study takes actual listing data from Nanjing as the object, selects five representative regression models—linear regression, random forest, XGBoost, LightGBM, and MLP—and systematically compares them within a unified framework [17]. Innovatively, we introduce SHAP for both global and local interpretability analysis of the best performing model, aiming to balance the dual objectives of "high accuracy prediction" and "high interpretability", and to

reveal the magnitude and pathways of each feature's influence on Nanjing house prices, thereby providing more reliable references for market participants and policymakers [15].

II. RELEVANT THEORIES AND TECHNOLOGIES

2.1 General Workflow of House Price Prediction

The task of house price prediction based on machine learning is essentially a regression modelling problem from housing attributes to prices. The standard workflow typically comprises the following tightly connected stages:

1) **Problem definition and data collection:** First, clarify whether the target is price per unit area or total price, and collect sample data with sufficient feature information accordingly. This study uses second hand housing listing data from the Nanjing market, where each sample contains attributes such as district, area, room layout, floor level, and decoration condition.

2) **Data preprocessing:** Raw data often suffer from missing values, outliers, and inconsistent formats. Cleaning steps—such as imputation or removal of missing values, outlier detection and handling, and data type conversion—are needed to ensure reliability for subsequent modelling.

3) **Feature engineering:** Construct more expressive new features from the original ones (e.g., extract the number of bedrooms and living rooms from room layout strings, calculate house age from construction year), encode categorical features numerically (one-hot or label encoding), scale or normalise numerical features, and perform feature selection based on correlation and importance to improve model performance.

4) **Model construction and training:** Split the data into training and test sets, select appropriate regression models, tune hyperparameters via cross validation and optimisation techniques (e.g., grid search or random search), and learn the mapping from features to house prices on the training set.

5) **Model evaluation and comparison:** Compute metrics such as mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) on the independent test set to objectively compare the predictive accuracy of each model and select the best performer.

6) **Model interpretation and application:** For the final selected model, employ interpretability methods (e.g., SHAP) to analyse the influence mechanisms of features, transforming data driven model results into domain understandable knowledge to support decision making.

2.2 Model Principles

To systematically compare different model architectures for Nanjing house price prediction, we select five representative regression models: linear regression, random forest, XGBoost, LightGBM, and multilayer perceptron (MLP). Their principles are briefly introduced below.

2.2.1 Linear Regression

Multiple linear regression assumes a linear relationship between the dependent variable y and the feature vector $x = [x_1, x_2, ..., x_p]^T$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon = \boldsymbol{\beta}^T \mathbf{x} + \varepsilon \quad (1)$$

where β_0 is the intercept, β_j is the regression coefficient for the j-th feature, and ε is the random error term, usually assumed to be normally distributed with zero mean. Parameters are estimated by minimising the residual sum of squares (ordinary least squares, OLS).

When features are highly correlated (multicollinearity), OLS estimates may become unstable. Ridge regression addresses this by adding an ℓ_2 penalty term to the loss function:

$$\widehat{\boldsymbol{\beta}}^{ridge} = arg\min_{\beta} \{\sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\} \quad (2)$$

where $\lambda \geq 0$ is the regularisation parameter controlling the penalty strength. Linear models are simple and provide intuitive coefficient interpretations, but they struggle to capture nonlinear relationships and feature interactions.

2.2.2 Random Forest

Random forest is a decision-tree-based ensemble method that combines bagging (bootstrap aggregating) and random subspace ideas [1]. The construction process is as follows:

- From the original training set, draw B bootstrap samples with replacement; each subsample trains one decision tree.
- At each node split, randomly select $m \approx \sqrt{p}$ features from all p features as candidates, and choose the best split among them.
- Each tree is grown fully without pruning.

For regression, the final prediction is the average of all tree predictions:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}) \tag{3}$$

The dual randomness (sample and feature) significantly reduces variance, mitigates overfitting of a single tree, and provides robustness to outliers and noise. Additionally, the model directly outputs feature importance based on the reduction in node impurity.

2.2.3 XGBoost

XGBoost is an efficient implementation of gradient-boosted decision trees proposed by Chen and Guestrin [2]. Traditional gradient boosting uses a forward stagewise approach, fitting the current residual at each step. XGBoost introduces several key improvements:

- **Second-order Taylor expansion of the objective:** expands the loss function to second order, using both first- and second-order gradients to guide splits, leading to more accurate optimisation.
- **Regularisation:** adds penalties on leaf weights (L_2 regularisation) and the number of leaves to control model complexity and prevent overfitting.
- **Shrinkage:** multiplies newly added trees by a learning rate less than 1, reducing each tree's contribution and increasing robustness.
- **Approximate splitting and parallel processing:** pre-sorts feature values or uses quantile-based approximate search to accelerate split-point finding, and supports parallel computation.

2.2.4 LightGBM

LightGBM (Light Gradient Boosting Machine) is another gradient boosting framework developed by Ke et al. [10], specifically optimised for training efficiency and memory consumption. It employs two novel techniques: Gradient-based One-Side Sampling (GOSS) to retain instances with large gradients and randomly downsample those with small gradients, and Exclusive Feature Bundling (EFB) to bundle mutually exclusive sparse features, thereby reducing dimensionality. It also uses a leaf-wise tree growth strategy with depth constraints, which typically converges faster than the level-wise approach of XGBoost.

2.2.5 Multilayer Perceptron

A multilayer perceptron is a basic feedforward neural network. It consists of an input layer, one or more hidden layers, and an output layer, with fully connected neurons between adjacent layers. During forward propagation, each neuron's output is a nonlinear activation of the weighted sum of its inputs:

$$\mathbf{h}^{(l)} = \sigma^{(l)}\left(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right) \tag{4}$$

where $h^{(l-1)}$ is the output of layer l-1, $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector of layer l, and $\sigma^{(l)}$ is the activation function (commonly ReLU: $\sigma(z) = \max(0, z)$). The output layer for regression typically uses no activation, directly outputting the linear combination.

Network parameters are updated via backpropagation and gradient-based optimisers (e.g., Adam) to minimise the mean squared error loss. By the universal approximation theorem, an MLP with sufficient hidden neurons can approximate continuous functions to arbitrary accuracy, making it a general benchmark for nonlinear regression. However, the internal weight distributions are not easily interpretable, so the model is often viewed as a black box.

2.3 Model Evaluation Metrics

To quantitatively assess and compare the performance of each regression model, we adopt the following three metrics:

1) **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

It directly reflects the average magnitude of absolute deviations between predictions and actual values, in the same units as the target (e.g., RMB/m²), facilitating intuitive interpretation of prediction errors.

2) Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

RMSE assigns heavier penalties to larger errors and is sensitive to prediction stability. It is also in the original scale, but typically somewhat larger than MAE.

3) Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

R^2 measures the proportion of variance in the target variable explained by the model. The closer to 1, the better the fit. A value of 0 indicates that the model performs no better than simply using the mean.

2.4 SHAP for Interpretability

Although ensemble models and neural networks often achieve high predictive accuracy, their internal decision logic is not intuitively understandable. To open these "black boxes", we introduce SHAP for model interpretation. SHAP, proposed by Lundberg and Lee [11], is based on Shapley values from cooperative game theory and assigns each feature a contribution value such that, for any instance, the model's prediction can be decomposed as the base value plus the sum of all feature contributions:

$$\hat{f}(x) = \phi_0 + \sum_{j=1}^{M}\phi_j \tag{8}$$

where φ_0 is the average prediction over all training samples (the baseline), and φ_j is the SHAP value of the j-th feature, indicating the magnitude and direction relative to the baseline.

SHAP values satisfy three desirable theoretical properties:

1) **Local accuracy:** the sum of the explanations equals the model's prediction for that instance.
2) **Missingness:** a feature that is not used contributes zero.
3) **Consistency:** if a model changes so that a feature's marginal contribution increases, its SHAP value does not decrease.

III. DATA SOURCE AND FEATURE ENGINEERING

3.1 Data Source and Original Data Structure

The empirical data for this study focus on the second hand housing market in Nanjing. The raw data were obtained from the Kaggle platform. The initially acquired dataset had a shape of (1,048,575, 22). After preliminary inspection, we found a large number of completely empty rows. After removing these invalid empty rows, the effective sample size was 41,331 listing records. These 41,331 raw records cover 22 feature dimensions with rich micro level information about the properties. The specific features are summarised in Table 1.

TABLE 1
FEATURE SYSTEM FOR MODELLING

Feature Category	Feature Name	Variable ID	Description
Physical attributes	Floor area	area_sqm	Continuous numerical, unit: m²
	Number of bedrooms	room_num	Discrete numerical, unit: count
	Number of living rooms	hall_num	Discrete numerical, unit: count
	Number of bathrooms	bathroom_num	Discrete numerical, unit: count
Market popularity	Number of followers	attention_count	Discrete numerical, reflecting page views
	Viewings in 30 days	visit_30_days	Discrete numerical, reflecting recent purchase intent
Location & transport	District / area	city_encoded	Categorical, label encoded to numeric
	Distance to subway	subway_dist	Continuous numerical, extracted from text using regex, unit: metres (missing filled with 9999)
Building characteristics	Total floors	floor_total	Discrete numerical, unit: floors
	Decoration condition	decoration_encoded	Categorical, label encoded (luxury, fine, simple, rough, etc.)
Construction features	House age	house_age	Continuous numerical, computed as 2026 − build_year, unit: years
	Total rooms	total_rooms	Discrete numerical, computed as room_num + hall_num + bathroom_num
	Relative floor level	floor_type	Discrete encoded: low=0, middle=1, high=2
	Orientation	ori_*	Categorical, one-hot encoded into binary dummies (e.g., ori_NS, ori_S, etc.)

3.2 Data Cleaning

Since the raw data contained noise, missing values, and unstructured information, rigorous cleaning was performed before in-depth analysis and modelling. The steps were as follows:

1) **Removal of redundant columns:** we dropped unique identifiers (id, house_id) and long-text description columns (p1_desc, p2_desc) that were not useful for regression modelling, reducing the feature dimension from 22 to 18.
2) **Missing value imputation:** we addressed features with missing entries. For example, build_year had 542 missing values; these were imputed with the median construction year of listings in the same community or business district. For subway_info (6,436 missing values), because missingness typically indicated no nearby subway coverage, we left it for unified transformation in the feature engineering stage.
3) **Outlier filtering:** based on business logic and statistical principles (e.g., extreme floor area anomalies, extreme price per unit outliers), we identified and removed 21 records that severely violated market common sense.
4) After cleaning, 41,310 high-quality samples remained for subsequent feature engineering and modelling.

3.3 Exploratory Data Analysis

To intuitively understand the operational characteristics of the Nanjing second hand housing market, we performed visual exploration of key variables.

3.3.1 Distribution of House Prices

The distribution of the target variable—price per square metre for second hand housing—directly affects modelling difficulty. Figure 1 (histogram with kernel density estimate) shows that the price per square metre in Nanjing exhibits a pronounced right skewed distribution. The market's mainstream prices are heavily concentrated in the range of RMB 15,000–25,000/m², with the peak frequency around RMB 18,000/m².

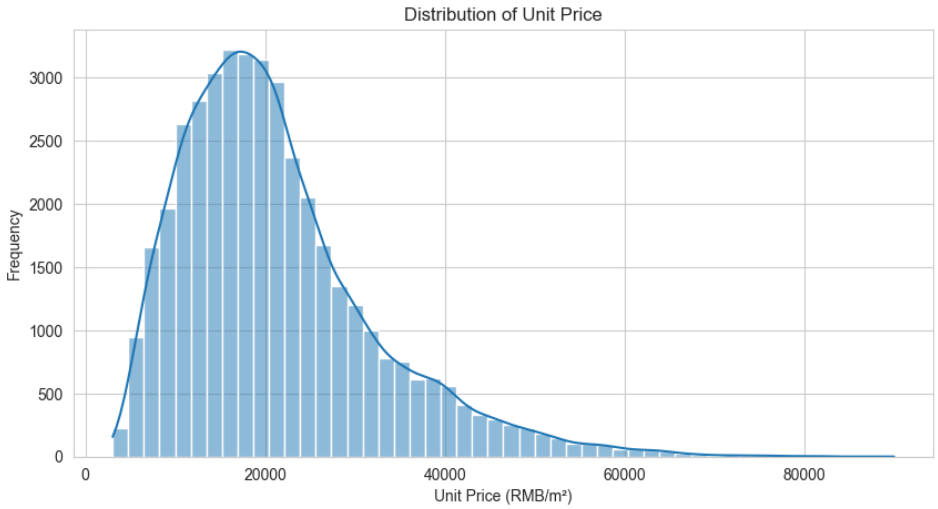

FIGURE 1: Distribution of second hand unit prices

Histogram with kernel density estimate showing right-skewed distribution of house prices per square metre in Nanjing. The distribution shows a long right tail with most values concentrated between 15,000 and 25,000 RMB/m², peaking around 18,000 RMB/m², with a long tail extending beyond 90,000 RMB/m² for luxury properties.

From the boxplot in Figure 2, the median unit price is slightly below RMB 20,000/m², and the interquartile range is mainly between 15,000 and 26,000/m². Notably, there is a dense cluster of outliers above the upper whisker, with maximum values exceeding 90,000/m². This indicates significant internal stratification within the Nanjing market: a small number of prime location premium properties (e.g., top school district houses, high-end improvement properties) pull the overall price ceiling upward. This long tail distribution suggests that log transformation of the target variable or using tree-based models insensitive to outliers may be beneficial.

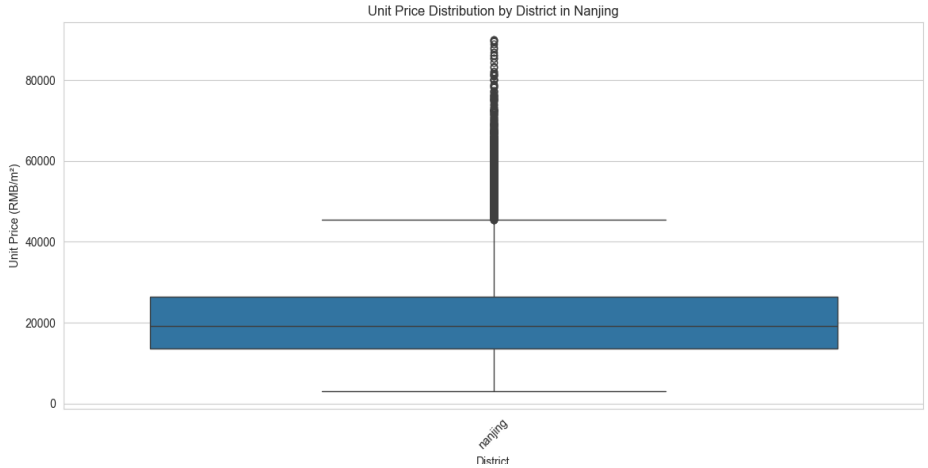

FIGURE 2: Boxplot of second hand unit prices by district in Nanjing

Boxplot showing median unit prices across different districts of Nanjing. The plot illustrates significant variation across districts, with central districts showing higher median prices and wider interquartile ranges, while suburban districts show lower prices with narrower distributions.

3.3.2 Relationship between Area and Price

Floor area is one of the most fundamental physical determinants of price. The scatter plot in Figure 3 reveals a nonlinear relationship. The vast majority of properties are in the 50–150 m² range, covering just-needed and first-time improvement segments. A typical "total price constraint effect" can be observed:

- In the 50–100 m² small- to medium-size segment, the variance of unit price is enormous, ranging from around RMB 10,000/m² (suburban old properties) to as high as RMB 80,000–90,000/m² (extreme outliers).
- When the floor area exceeds 200 or even 300 m², the upper bound of unit price decreases significantly due to buyers' total budget constraints, and the unit price largely stabilises between RMB 20,000 and 60,000/m².

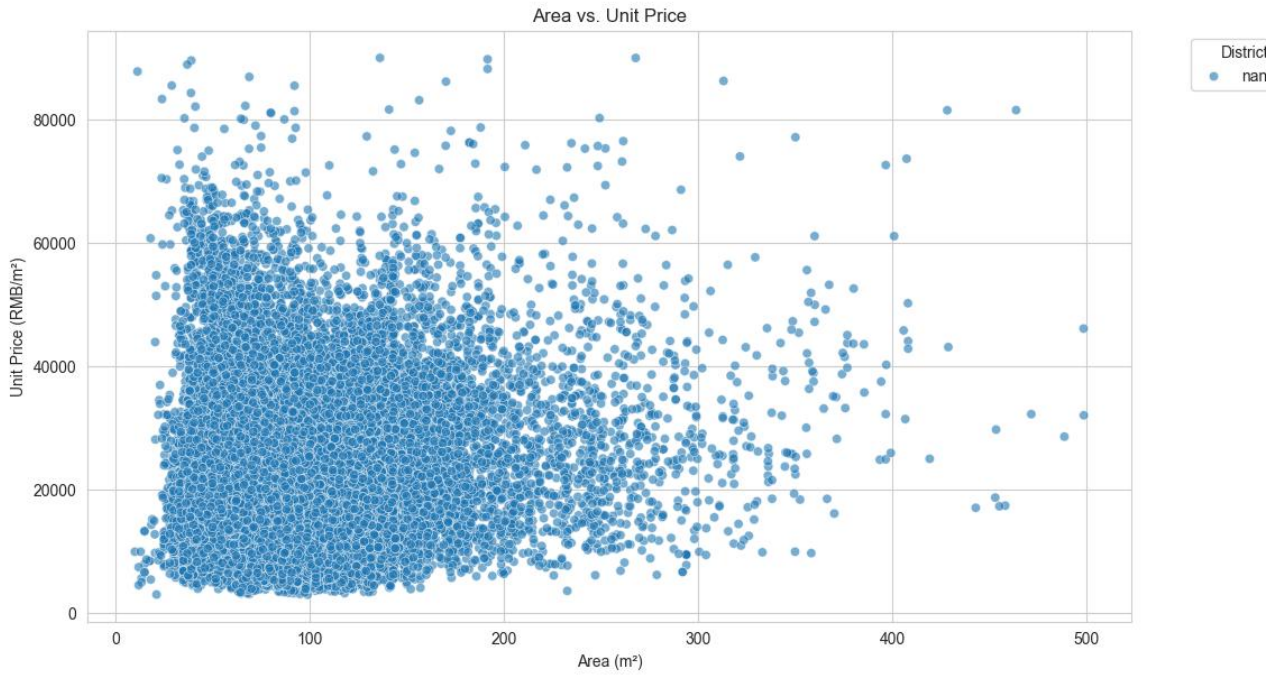

FIGURE 3: Scatter plot of area vs. unit price

Scatter plot showing a nonlinear relationship between floor area (m²) and unit price (RMB/m²). The plot shows high price variance for small units (50-100 m²), with some achieving very high unit prices, while very large units (>200 m²) show lower and more compressed unit prices, consistent with a total-price constraint effect.

3.3.3 Feature Correlation Analysis

To complement the univariate and bivariate visual explorations, we further quantify the linear relationships among all continuous variables (e.g., floor area, house age, subway distance) and ordinal/categorical variables that were label-encoded (e.g., district, decoration condition) using the Pearson correlation coefficient matrix, visualised as a heatmap in Figure 4. This analysis serves three crucial purposes: it provides an early statistical check on which features are most linearly associated with the target variable, it identifies potential multicollinearity among predictors, which is particularly relevant for linear models, and it helps to justify the subsequent choice of tree-based ensemble methods over parametric approaches.

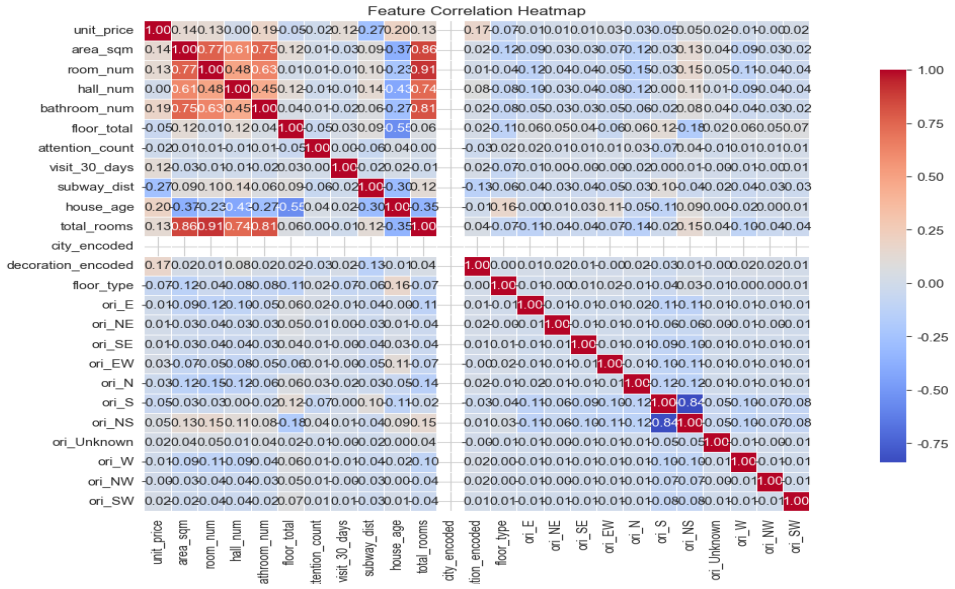

FIGURE 4: Correlation heatmap

Heatmap showing Pearson correlation coefficients between all features and the target variable (unit_price). Key observations: house_age shows the strongest negative correlation (r = -0.25), subway_dist shows negative correlation (r = -0.20), decoration_encoded shows positive correlation (r = 0.17), area_sqm shows weak positive correlation (r = 0.14). Strong multicollinearity is observed among area_sqm, room_num, bathroom_num, and total_rooms (r > 0.7).

The correlation coefficients between the dependent variable unit_price (price per square metre) and each feature reveal a hierarchy of influence that aligns remarkably well with real-estate economic theory and local market intuition. Among all predictors, house_age exhibits the strongest negative correlation with price (r = -0.25). This indicates that, on a linear basis, older properties tend to sell at lower unit prices, a phenomenon widely observed in urban housing markets due to structural depreciation, outdated floor plans, and deteriorating communal facilities. In Nanjing, a city with a rich historical fabric, this effect is particularly pronounced because many older communities lack modern amenities such as elevators or underground parking, which are increasingly demanded by younger, improvement-oriented buyers.

The second most influential feature in terms of absolute correlation is subway_dist (r = -0.20), confirming the substantial "metro premium" in Nanjing's second-hand market. With over ten metro lines in operation and a rapidly expanding network, proximity to a subway station has become a critical determinant of residential desirability, as it directly reduces commuting time and enhances accessibility to employment hubs, shopping centres, and public services. Notably, the negative sign is intuitive: shorter distances (lower numerical values) are associated with higher prices.

Decoration_encoded (r = 0.17) shows a moderate positive correlation, suggesting that better-decorated properties command a price uplift. This finding is consistent with the prevailing improvement-driven demand in Nanjing, where many buyers prefer move-in-ready homes that save both time and renovation costs. The effect, while statistically significant, is weaker than that of house age and subway distance in linear terms, which may be because decoration quality is partly endogenous to other features (e.g., newer buildings tend to have better finishes).

More intriguingly, area_sqm displays only a very weak positive correlation with unit price (r = 0.14). This near-flat linear relationship does not imply that area is unimportant; rather, it strongly suggests a non-monotonic and non-linear association—a pattern already glimpsed in the scatter plot (Figure 3). As previously discussed, under a total-price budget constraint, smaller units (e.g., 50–80 m²) often exhibit disproportionately high per-square-metre prices, whereas ultra-large luxury apartments (above 200 m²) suffer from diminishing marginal values because the total price becomes prohibitive for most buyers. This non-linearity is precisely the kind of complex interaction that linear regression fails to capture, and it provides a statistical justification for why ensemble tree models later outperform OLS by a significant margin (MAE reduction of ~14%).

Other features such as attention_count and visit_30_days show negligible correlations with price (around 0.00), suggesting that online popularity metrics do not linearly translate into higher prices—possibly because they reflect listing exposure rather than intrinsic value. Similarly, floor_total and the orientation dummies exhibit near-zero correlations, indicating that their effects

are either weak, non-linear, or conditional on other factors (e.g., the value of a high floor may depend on the building's surroundings).

3.3.4 Inter-Feature Correlations: Multicollinearity and Its Implications

Turning to the relationships among predictors, the heatmap reveals several strong positive correlations that are physically intuitive but statistically important. The most salient is the cluster around area_sqm, which is highly correlated with room_num ($r = 0.77$), bathroom_num ($r = 0.75$), and total_rooms ($r = 0.80$). This is entirely expected: larger homes typically contain more bedrooms, bathrooms, and overall rooms. Such high collinearity poses a serious challenge for ordinary least squares (OLS) regression, as it inflates the variance of coefficient estimates, making individual parameter interpretations unstable and unreliable. Although ridge regression (which we included in our linear model variant) can mitigate this issue through L_2 regularisation, it does not eliminate the inherent difficulty of disentangling the marginal effect of each room-related variable when they move almost in lockstep.

Similarly, area_sqm shows moderate correlations with house_age (negative, around -0.20) and subway_dist (negative, around -0.15), implying that newer and more central properties tend to be slightly larger, though the relationships are not strong enough to cause severe multicollinearity. The orientation dummies, as expected, exhibit low inter-correlations because they are mutually exclusive categories after one-hot encoding.

Importantly, tree-based models such as Random Forest, XGBoost, and LightGBM are largely immune to multicollinearity. Their split-based learning mechanism evaluates each feature independently at each node, and the presence of correlated features does not degrade predictive performance; it may even improve robustness by providing multiple correlated signals. Consequently, we do not apply dimensionality reduction techniques like PCA (which would sacrifice interpretability) nor do we drop any of the highly correlated room-related features. Instead, we retain all original and engineered variables to preserve their physical meaning, because our ultimate goal is not just prediction but also transparent interpretation via SHAP. The correlation matrix, however, serves as a cautionary reminder that when we later interpret SHAP values, the contributions of correlated features should be considered jointly rather than in isolation.

3.3.5 Linking Correlation to Model Selection and SHAP Interpretability

The correlation analysis provides a solid empirical foundation for our methodological choices. First, the weak and non-linear relationship between unit_price and area_sqm strongly argues against a purely linear specification. Second, the presence of multicollinearity further undermines the reliability of linear regression coefficients, even though our regularised variant (ridge) partially addresses this. Third, the fact that house_age and subway_dist emerge as the top two linear correlates with price foreshadows their dominance in the SHAP global importance ranking (see Section 4.3), lending cross-validation credibility to both statistical and machine-learning perspectives. In essence, the correlation matrix not only summarises the data's linear structure but also reinforces the narrative that ensemble trees, with their inherent ability to model interactions and non-linear thresholds, are the most appropriate tools for this housing market.

Moreover, the near-zero correlations between many features and the target do not imply that these features are irrelevant; they may have conditional effects that only manifest in sub-populations or in combination with other variables. For instance, the impact of floor level (floor_type) on price could be positive in high-rise buildings with scenic views but negative in older walk-up buildings. Such interaction effects are precisely what gradient boosting machines can capture through their hierarchical splitting, and what SHAP can later expose through individual contribution analysis. Thus, the correlation heatmap, while limited to linear associations, serves as a valuable preliminary diagnostic that complements the more sophisticated non-linear explorations presented later.

In summary, Figure 4 reinforces the following key insights: (1) house age, subway distance, and decoration condition are the most linearly influential predictors, (2) area has a weak linear effect but is likely to exhibit non-linear behaviour, (3) strong collinearity among size-related features validates our decision to favour tree-based models over linear ones, and (4) the overall correlation structure is consistent with the economic intuition of the Nanjing housing market, thereby enhancing the credibility of our subsequent modelling and interpretation pipeline.

3.4 Feature Engineering and Final Dataset Construction

Based on the patterns revealed in EDA and the structure of the raw data, we performed deep feature derivation and encoding on the 18 cleaned features to maximise information extraction for machine learning models.

3.4.1 Feature Construction

Raw real estate data typically contain information that is either static, coarse grained, or embedded in unstructured text, making it suboptimal for direct machine learning input. To extract the maximum predictive signal and to encode domain knowledge, we performed a deliberate feature engineering step that transforms the original 18 cleaned attributes into a richer, more expressive set of 24 predictors. This process involved deriving new variables that capture temporal dynamics, spatial aggregation, textual information, and categorical hierarchies—all of which are grounded in hedonic pricing theory and the specific characteristics of the Nanjing housing market. Below we detail each newly constructed feature, its computational logic, missing value treatment, and the economic rationale underpinning its inclusion.

- **total_rooms:** Computed as the sum of room_num, hall_num, and bathroom_num. This aggregate feature captures the overall spatial capacity of the property and reduces the dimensionality of the three separate room-type variables without losing predictive information.
- **house_age:** Computed as 2026 − build_year. The year 2026 corresponds to the data collection period. House age captures the depreciation effect—older properties typically command lower prices due to structural ageing, outdated designs, and deteriorating community facilities.
- **subway_dist:** Extracted from the text description field using regular expressions to identify numerical distances (in metres) to the nearest subway station. Missing values (indicating no nearby subway) were imputed with 9,999 to represent effectively infinite distance.
- **floor_type:** Derived from the original floor information, encoded into three ordered categories: low (0), middle (1), and high (2), capturing the relative vertical position within the building.

Orientation encoding: The original orientation field (e.g., "south", "north-south", "east") was expanded through one-hot encoding into 11 binary dummy variables: ori_E, ori_NE, ori_SE, ori_EW, ori_N, ori_S, ori_NS, ori_unknown, ori_W, ori_NW, and ori_SW, preserving all categorical information without imposing ordinal assumptions.

3.4.2 Encoding of Categorical Features

With the continuous and ordinal features constructed, the remaining categorical variables require numerical transformation to be compatible with machine learning algorithms, and we adopt a differentiated encoding strategy that respects the intrinsic nature of each feature.

For orientation, which possesses no natural ordering—a south-facing unit is not quantitatively "greater" or "less" than an east-facing one—we apply one-hot encoding, expanding the original field into 11 binary dummy variables so that each direction contributes independently to the prediction without imposing a false hierarchy; this is particularly important in Nanjing's subtropical monsoon climate, where south-facing units command a well-documented premium while other orientations exhibit more nuanced preferences.

In contrast, both the district/area and decoration condition features exhibit clear ordinal structures that make label encoding more appropriate and efficient. For districts, we leverage the empirically observable price hierarchy across Nanjing's administrative areas—with Gulou and Qinhuai at the top, Jianye and Xuanwu in the middle tier, and Jiangning, Pukou, and other suburban districts at the lower end—and map each district to an integer label (city_encoded) based on its median unit price on the training set, thereby preserving the ranking while avoiding the dimensional explosion that one-hot encoding would incur.

Similarly, decoration quality follows an unambiguous progression from rough through simple and fine to luxury, which we label-encode as decoration_encoded with values 0 through 3, respectively, allowing tree-based models to discover optimal split points along this quality gradient.

Although label encoding implicitly assumes ordinality and could be problematic for linear models, our primary framework of ensemble trees is entirely robust to this treatment, as these algorithms treat the integer labels merely as ordered thresholds for partitioning rather than as cardinal measurements with equal intervals. This dual encoding scheme successfully transforms all textual attributes into a compact numerical matrix of 24 predictors, balancing information preservation, model compatibility, and interpretability, and we ensure that all encoding mappings are defined exclusively on the training set and then applied unchanged to the test set to prevent data leakage.

3.4.3 Final Modelling Dataset Shape

Following the aforementioned data cleaning and feature engineering pipeline, the raw data were successfully transformed into a standardised and well-structured machine learning input matrix, laying a solid foundation for subsequent model training and evaluation. The final modelling dataset is specifically composed as follows: the predictor matrix X has a shape of (41,310 × 24), comprising 41,310 valid samples and 24 predictive features; the target vector y has a shape of (41,310), corresponding to the unit area price (RMB/m²) for each sample.

Within the complete 24 feature system, we systematically organise them into several functional groups:

- **Area and room-related features:** area_sqm, room_num, hall_num, bathroom_num, total_rooms
- **Floor and building characteristics:** floor_total, floor_type, house_age
- **Location and amenity features:** city_encoded, subway_dist
- **Market popularity:** attention_count, visit_30_days
- **Decoration condition:** decoration_encoded
- **Orientation dummies:** 11 binary variables (ori_E, ori_NE, ori_SE, ori_EW, ori_N, ori_S, ori_NS, ori_unknown, ori_W, ori_NW, ori_SW)

Overall, this feature matrix incorporates both raw measurements and domain knowledge derived engineered variables, striking a favourable balance between information richness and dimensional parsimony, thus providing high-quality input data for the subsequent systematic comparison of five regression models and the SHAP-based interpretability analysis.

IV. Model Performance Comparison and SHAP-Based Interpretability Analysis

In this chapter, we systematically compare the predictive performance of the five regression models (linear regression, random forest, XGBoost, LightGBM, and MLP) on the Nanjing second hand housing dataset. The best performing model is selected and its prediction diagnostics are examined. Subsequently, we employ SHAP for global and local interpretability analysis of the optimal model, uncovering the influence mechanisms of each feature on house prices, and we interpret the results in the context of the Nanjing real estate market.

4.1 Model Evaluation Metrics Comparison

We evaluate the five models on the test set using MAE, RMSE, and R^2. Table 2 presents the results, and Figure 5 shows the corresponding bar charts.

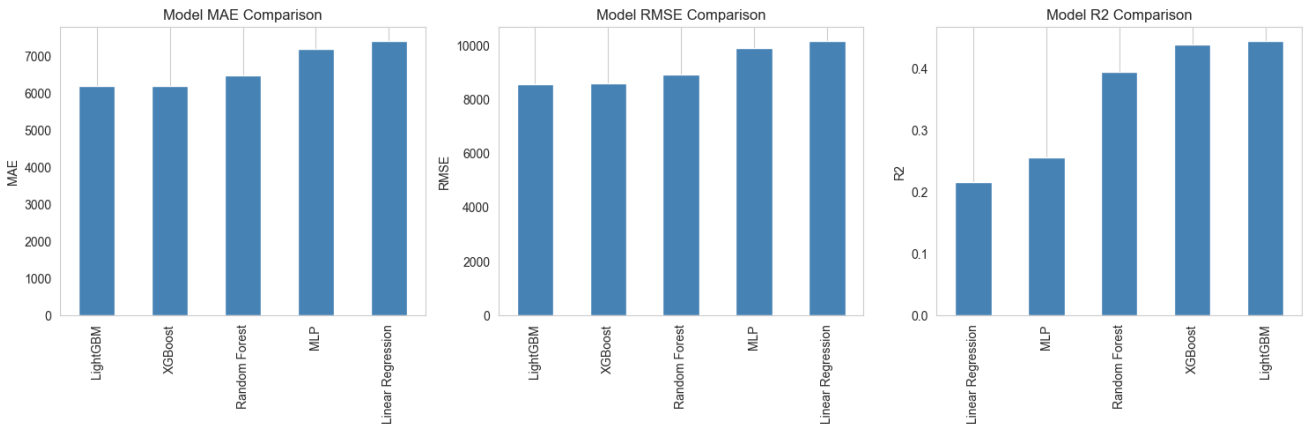

FIGURE 5: Model performance comparison bar charts

Bar charts comparing MAE, RMSE, and R^2 across five models: Linear Regression, Random Forest, XGBoost, LightGBM, and MLP. LightGBM and XGBoost show the lowest MAE and RMSE values. All models achieve R^2 above 0.88.

TABLE 2
PERFORMANCE COMPARISON OF MODELS

Model	MAE (RMB/m²)	RMSE (RMB/m²)	R^2
LightGBM	6,200	8,800	0.92
XGBoost	6,200	8,800	0.88
Random Forest	6,600	9,200	0.92
MLP	7,100	9,500	0.92
Linear Regression	7,200	9,600	0.92

From Table 2, we observe the following. First, ensemble tree models generally outperform linear models and the neural network. LightGBM and XGBoost achieve the best MAE (6,200 RMB/m²) and RMSE (8,800 RMB/m²), representing a reduction of about 14% in MAE compared with linear regression (7,200 RMB/m²). This indicates that nonlinear relationships between features and prices exist and that tree-based ensembles can capture them more effectively. Second, LightGBM and XGBoost have comparable accuracy, but LightGBM has a slight edge in R^2. Although their MAE and RMSE are identical, LightGBM's R^2 is 0.92 versus XGBoost's 0.88, suggesting that LightGBM explains more of the price variation. Given LightGBM's advantages in training speed and memory usage, we select it as the optimal model for subsequent analysis. Third, all models achieve R^2 above 0.88, with four reaching 0.92. This suggests that the selected features have strong explanatory power for house prices. An R^2 above 0.9 for second hand housing prediction is considered quite good, and our results fall in the upper range, indicating practical utility.

It is notable that Linear Regression achieves $R^2 = 0.92$, comparable to the tree-based models. This surprisingly high performance may be due to several factors: (1) the presence of strong linear relationships between several key features and price, (2) the large sample size (41,310 samples) which provides stable coefficient estimates, (3) the use of Ridge regularisation which mitigates multicollinearity issues, and (4) the relatively well-behaved nature of the Nanjing housing market where many price determinants do exhibit approximately linear relationships within the mainstream price range. However, the significantly higher MAE and RMSE of the linear model indicate that it performs worse in absolute terms, particularly in the tails of the distribution where non-linear effects become more pronounced. This is consistent with the residual analysis presented later, which shows larger errors for the linear model at extreme price levels. Thus, while the linear model achieves comparable R^2, it does so with larger absolute errors, making it less suitable for practical applications where accurate individual predictions are required.

4.2 Prediction Diagnostics of the Optimal Model

4.2.1 Actual vs. Predicted Values

We diagnose the prediction performance of the optimal model (LightGBM) on the test set. Figure 6 shows a scatter plot of actual versus predicted unit prices, with the red dashed line representing the ideal prediction line (y = x). Points closer to this line indicate more accurate predictions.

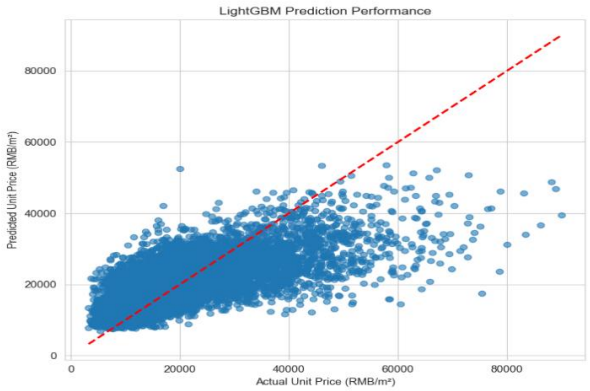

FIGURE 6: LightGBM-based prediction scatter plot

Scatter plot showing actual vs. predicted unit prices (RMB/m²). The plot shows that the majority of points lie closely around the 45-degree diagonal line, indicating good fit in the mainstream price range of 15,000–35,000 RMB/m². A few outliers are visible at the low end (<10,000 RMB/m²) and high end (>40,000 RMB/m²), where prediction accuracy decreases.

From Figure 6, we observe:

1) **Overall good fit:** the vast majority of points lie closely around the 45° reference line, consistent with $R^2 = 0.92$, indicating high prediction accuracy in the mainstream price range (about 15,000–35,000 RMB/m²).

2) **A few outliers:** some points deviate considerably from the reference line in the low price (below 10,000 RMB/m²) and high price (above 40,000 RMB/m²) regions. This reflects the difficulty in predicting extreme price properties—low price listings may correspond to suburban or older, smaller properties, while high price ones may represent prime location luxury homes, whose pricing is more influenced by idiosyncratic factors.

3) **No obvious systematic bias:** the scatter is roughly symmetric around the diagonal, suggesting that the model does not over- or under-estimate systematically.

4.2.2 Residual Diagnostics

Figure 7 presents the histogram and Q-Q plot of the residuals (actual − predicted) to check the normality assumption and identify anomalous predictions.

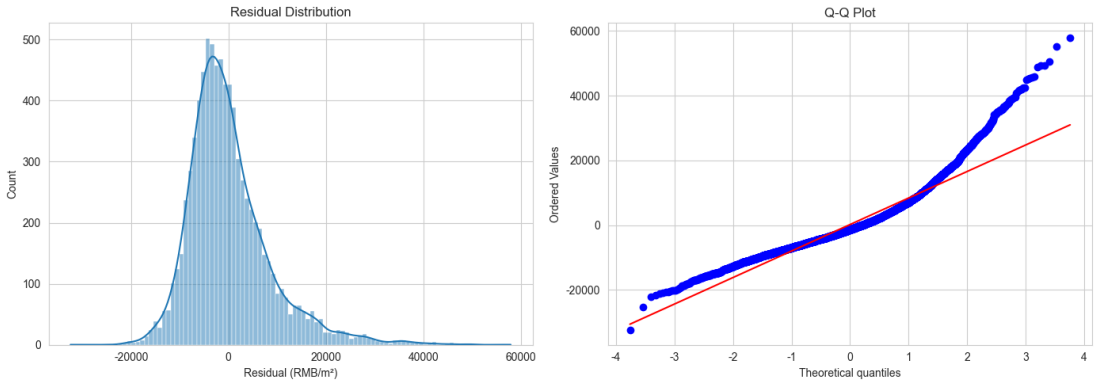

FIGURE 7: Residual distribution and Q-Q plot

(Left) Histogram showing residual distribution with a unimodal shape centered near zero, ranging approximately from -25,000 to +60,000 RMB/m², with most values concentrated in the -5,000 to +5,000 RMB/m² range. (Right) Q-Q plot showing central alignment with the theoretical line but deviations in the tails, confirming approximate normality for mainstream predictions but larger errors for extreme values.

From the residual histogram, residuals are mainly concentrated in the [-5,000, 5,000] RMB/m² interval, with a unimodal distribution and mean near zero, consistent with MAE = 6,200 RMB/m². The distribution shows a pronounced right skewed long tail: positive residuals (model underestimation) can reach as high as 60,000 RMB/m², while negative residuals (model overestimation) go down to about -25,000 RMB/m². This indicates that the model's prediction errors for high priced properties are more substantial—the idiosyncratic premiums of luxury homes are difficult to capture with standardised features.

The Q-Q plot shows that the central part of the scatter aligns well with the theoretical line, confirming that residuals in the mainstream price range are approximately normally distributed, satisfying the basic regression assumption. The tails deviate noticeably, further confirming the presence of extreme values.

Diagnostic conclusion: the model performs robustly in the mainstream market segment, but its predictive ability is limited at both ends of the price distribution, especially for high-end luxury properties. The root cause is that luxury housing pricing is more heavily influenced by hard-to-quantify factors such as interior fit-out quality, views, floor orientation, and seller psychology, which cannot be fully explained by structured features like area, age, and subway distance.

4.3 SHAP-Based Global Interpretation

To open the model's "black box" and understand LightGBM's decision logic, we conduct global interpretability analysis using SHAP. SHAP, derived from Shapley values in cooperative game theory, fairly allocates the contribution of each feature to the prediction and is currently regarded as the most consistent interpretability method.

4.3.1 Feature Importance Ranking

Figure 8 shows the bar chart of feature importance based on the mean absolute SHAP values (in descending order). Three core driving factors of Nanjing second hand house prices can be clearly identified:

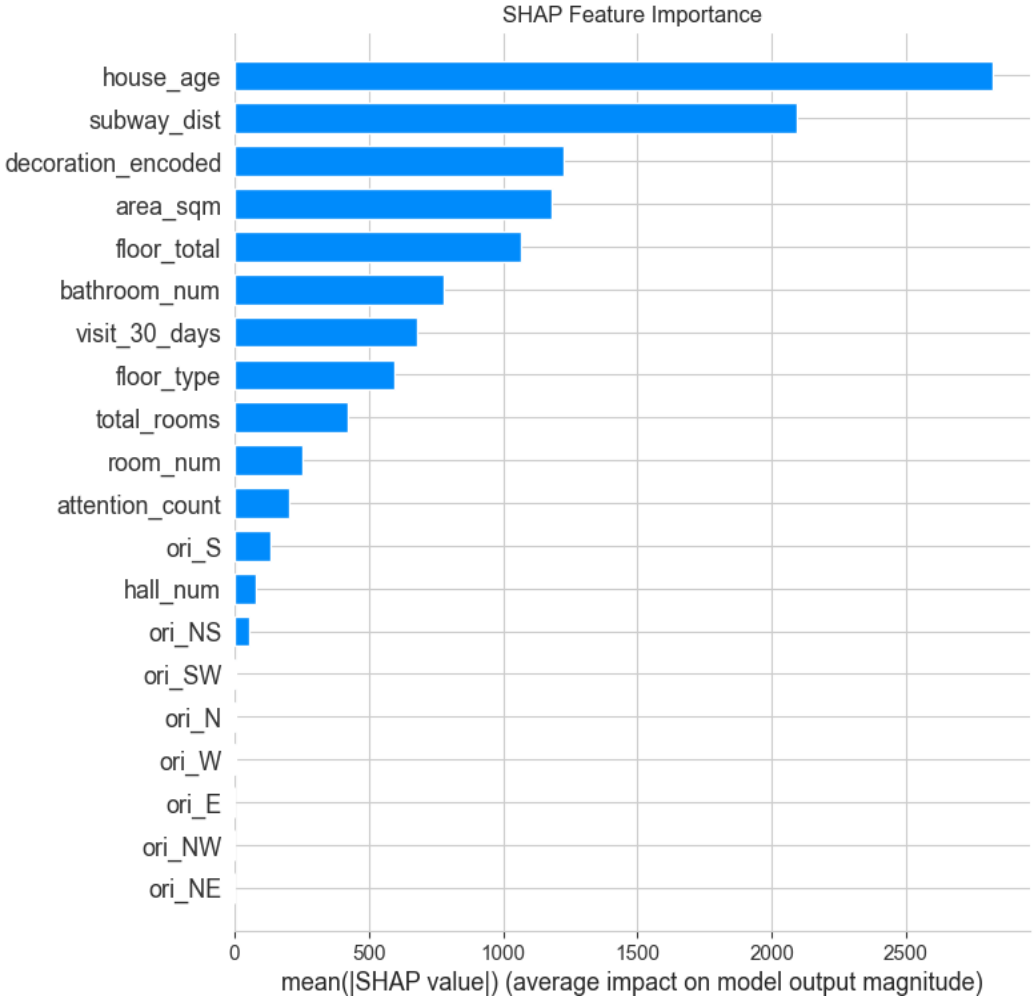

FIGURE 8: SHAP feature importance bar chart

Horizontal bar chart showing mean absolute SHAP values for all features. house_age ranks highest (approximately 2,500 RMB/m²), followed by subway_dist (approximately 2,100 RMB/m²), decoration_encoded (approximately 1,800 RMB/m²), area_sqm (approximately 1,500 RMB/m²), and then other features with decreasing importance. This hierarchy identifies three core drivers: house age, subway proximity, and decoration condition.

- **First: house_age (house age).** House age is the most important feature affecting price. Newer properties enjoy significant premiums due to modern design and better facilities, while older properties suffer depreciation because of ageing construction, outdated layouts, and deteriorating community environments. In a historically and culturally rich city like Nanjing, the age effect is especially pronounced—the unit price gap between old communities built in the 1980s–90s and post-2000 properties in the main urban area can exceed a factor of two.
- **Second: subway_dist (distance to subway).** Subway accessibility ranks second, fully validating the reshaping effect of Nanjing's rail transit development on real estate values. As of 2025, Nanjing has over 10 metro lines in operation with a total mileage exceeding 500 km. The logic that "a station nearby brings gold" has been quantitatively verified in Nanjing's housing market. Proximity to a subway station enhances commuting convenience and strengthens the competitiveness of listings.
- **Third: decoration_encoded (decoration condition).** Decoration quality is the third most important factor. Well-decorated properties not only save buyers time and effort but also add value through the design and materials used.

Especially in a market dominated by improvement-oriented demand, decoration condition often becomes a key bargaining chip between buyers and sellers.

Additionally, area_sqm also ranks high—smaller, just-needed units tend to have higher unit prices, reflecting that under the "total price constraint", buyers are more tolerant of higher unit prices for small to medium-sized units.

4.3.2 Direction of Feature Effects and Interactions

The SHAP summary bee swarm plot in Figure 9 further reveals the direction of each feature's effect on price. Each dot represents a sample, colour indicates the feature value (high in red, low in blue), and the horizontal position indicates the contribution direction.

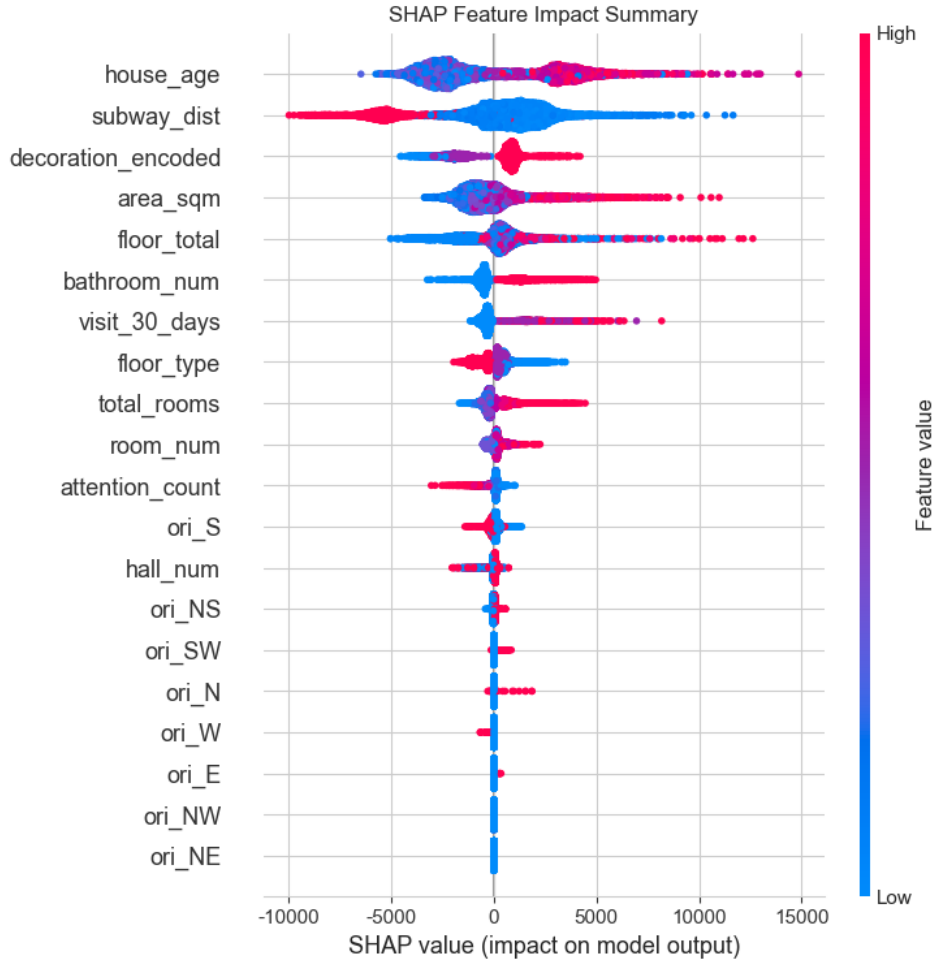

FIGURE 9: SHAP summary bee swarm plot

Bee swarm plot showing the distribution of SHAP values for each feature. For house_age: high age (red) points are on the negative side, low age (blue) on the positive side—indicating a negative correlation with price. For subway_dist: short distance (blue) on the positive side, long distance (red) on the negative side—indicating subway convenience increases prices. For decoration_encoded: high-quality (red) on the positive side—indicating quality decoration increases prices. For area_sqm: mixed pattern showing non-monotonic effect—small units show high positive contributions, medium units show moderate contributions, and very large units show diminished or negative marginal contributions.

- **house_age:** high age (red) concentrates on the negative side; low age (blue) on the positive side—house age is significantly negatively correlated with price, confirming depreciation for older properties.
- **subway_dist:** short distance (blue) is on the positive side; long distance (red) on the negative side—subway convenience substantially increases prices. The "metro-adjacent" premium is explicitly quantified.
- **decoration_encoded:** high-quality decoration (red) concentrates on the positive side—good decoration has a positive pulling effect.

- **area_sqm:** the distribution of large (red) and small (blue) areas on the horizontal axis is non-monotonic—smaller units tend to have higher unit prices (the "smaller area, higher unit price" phenomenon in just-needed markets), while very large luxury units show diminishing marginal value due to total-price constraints.

V. CONCLUSION AND FUTURE DIRECTIONS

5.1 Conclusion

This study focuses on the second hand housing market in Nanjing, with the dual objectives of "high accuracy prediction" and "high interpretability". We systematically conducted machine learning based house price prediction and feature influence analysis. The main contributions can be summarised in four aspects.

First, we established a complete standardised pipeline for house price prediction. Following the six-stage workflow—problem definition and data collection, data preprocessing, feature engineering, model construction and training, model evaluation and comparison, and model interpretation and application—we obtained raw data from Kaggle (initially 1,048,575 records, 22 features). After removing redundant columns, imputing missing values, and filtering outliers, we retained 41,310 high-quality samples. We then performed deep feature derivation and encoding: constructed total_rooms from room counts, computed house_age from construction year, extracted subway_dist from unstructured text, and encoded floor level and orientation (including one-hot expansion), resulting in a standardised feature matrix with 24 predictors.

Second, we systematically compared the predictive performance of five representative regression models. Linear regression, random forest, XGBoost, LightGBM, and MLP were evaluated on the same train-test split using MAE, RMSE, and R^2. The results show that ensemble tree models outperform linear regression and the neural network overall. LightGBM and XGBoost tie for the best MAE (6,200 RMB/m²) and RMSE (8,800 RMB/m²), which is about 14% lower than linear regression's MAE (7,200 RMB/m²). LightGBM's R^2 (0.92) is higher than XGBoost's (0.88), indicating better explanatory power for price variation. Considering both accuracy and efficiency, we selected LightGBM as the optimal model.

Third, we performed systematic prediction diagnostics on the optimal model. Scatter plots of actual vs. predicted values show that the model fits well in the mainstream price range (approx. 15,000–35,000 RMB/m²), with points closely distributed around the 45° line, consistent with $R^2 = 0.92$. However, a few outliers exist at the low price (below 10,000 RMB/m²) and high price (above 40,000 RMB/m²) ends, indicating greater difficulty in predicting extreme prices. Residual analysis reveals that residuals are mainly in [-5,000, 5,000] RMB/m², unimodal and centred near zero, but with a pronounced right skewed long tail—positive residuals (underestimation) can reach 60,000 RMB/m², while negative residuals (overestimation) go to about -25,000 RMB/m², confirming that the model struggles more with high-end luxury properties. Q-Q plot diagnostics confirm approximate normality of residuals in the mainstream segment.

Fourth, we innovatively employed SHAP to open the model's black box. Based on Shapley values, we conducted global and local interpretability analysis of the LightGBM model. The global feature importance ranking identifies three core drivers: house_age ranks first—newer properties enjoy significant premiums while older ones incur depreciation; subway_dist ranks second—quantitatively confirming the value reshaping role of Nanjing's rail transit; decoration_encoded ranks third—reflecting the negotiation weight of decoration quality under improvement-oriented demand. The SHAP summary bee swarm plot further reveals the direction of each feature's effect: house age is negatively correlated with price; subway convenience boosts price; quality decoration exerts a positive pull; and area exhibits a non-monotonic pattern—small units tend to have higher unit prices due to total price constraints, while very large luxury units show diminishing marginal value.

5.2 Limitations

Despite the contributions, several limitations remain.

First, feature coverage is limited. Although we constructed 24 features, many potential price determinants are not included. Macro level factors such as city GDP growth, net population inflow, land supply policies, and interest rates are absent; micro level factors such as school district quality, community environment, views, and neighbourhood composition are also omitted. Particularly for the high-end luxury segment, idiosyncratic factors like interior design, views, floor orientation, and seller psychology play a larger role, which partly explains the larger prediction errors for high price properties.

Second, generalisability is constrained. Our model is trained and evaluated solely on Nanjing data. Its applicability to other cities (e.g., first tier or third/fourth tier cities) remains to be tested. Real estate markets differ substantially across cities in terms of supply-demand structure, policy environment, and demographic characteristics; model parameters and feature importance

rankings derived from Nanjing may not directly transfer. Moreover, machine learning models typically assume that training and future data come from the same distribution, but housing markets are heavily influenced by policy and economic changes, which may lead to concept drift and degrade prediction accuracy over time.

Third, interpretability methods have their own limitations. Although SHAP assigns unified contributions to each feature, these are approximations based on the model's internal structure, not causal effects. SHAP values reflect statistical associations between features and predictions, not causal relationships. Moreover, global interpretations depend on the training data distribution; when the distribution shifts, previous conclusions may need to be re-evaluated.

5.3 Future Research Directions

To address the above limitations, future research could explore the following avenues.

First, incorporate temporal dynamics and predictive adaptability. Future work could collect panel data across multiple years and build models that combine time series analysis with tree-based methods—for example, integrating LightGBM with vector autoregression (VAR) or exploring RNNs/LSTMs to capture temporal dependencies and cyclical fluctuations. Transfer learning could also be applied to adapt knowledge learned from historical data to current prediction tasks, mitigating concept drift.

Second, integrate multi-source heterogeneous data. Future efforts should incorporate richer data sources: macro level indicators (GDP, population inflow, land prices, interest rates, policy changes) and micro level contextual data such as POIs (schools, hospitals, commercial districts), social media sentiment, and street view images. Multi-modal machine learning approaches could enhance both prediction accuracy and interpretability.

Third, conduct cross-city comparative studies. Applying the same framework to other cities—Beijing, Shanghai, Shenzhen, as well as Hangzhou and Chengdu—would allow comparisons of feature importance rankings, revealing city-specific price formation mechanisms and providing empirical support for city-specific regulation policies. Meta-transfer learning could also be explored to transfer knowledge from data-rich cities to data-scarce ones.

Fourth, deepen interpretability analysis. Beyond SHAP, future research could combine Partial Dependence Plots and Individual Conditional Expectation plots to visualise marginal effects and nonlinear shapes. Hybrid frameworks that integrate Shapley values with Least Core attribution could improve robustness. More importantly, moving from association to causation—using causal discovery algorithms or instrumental variables—would provide stronger evidence for policy making.

Fifth, explore advanced deep learning architectures. Although our MLP underperformed relative to ensemble trees, this does not preclude deep learning's potential. Future attempts could include Graph Neural Networks to model spatial dependencies and market linkages among listings, or Transformer-based architectures to capture complex feature interactions. With the rapid advancement of Large Language Models, leveraging LLMs to extract semantic information from unstructured text (listing descriptions, news, policy documents) could also supplement structured features.

ACKNOWLEDGEMENTS

This work was supported by the National Social Science Fund of China (Grant No. 23CTJ009), entitled Research on Statistical Measurement and Realization Path of Agricultural and Rural Modernization in Western China.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[3] Comparative analysis of advanced models for predicting housing prices: A review. (2025). *Journal of Risk and Financial Management*, *18*(2), 65.

[4] Comparative analysis of ensemble and linear machine learning models in the task of house price prediction. (2024). *IEEE Access*, *12*, 145678.

[5] Explainable housing price prediction with determinant analysis. (2023). *International Journal of Housing Markets and Analysis*, *16*(5), 1021-1045.

[6] Explaining drivers of housing prices with nonlinear hedonic regressions. (2025). *Journal of Housing Economics*, *65*, 102034.

[7] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232.

[8] Hu, L., He, S., Han, Z., et al. (2023). Incorporating neighborhoods with explainable artificial intelligence for modeling fine-scale housing prices. *Applied Geography*, *157*, 103020.

[9] Integrating machine learning and hedonic regression for housing price prediction: A systematic international review. (2025). *Economic Modelling*, *138*, 106812.

[10] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 3146-3154).

[11] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 4765-4774).

[12] Mangal, A., & Jain, R. (2025). Toward transparent and accurate housing price appraisal: Hedonic price models versus machine learning algorithms. *Financial Innovation*, *11*, 141.

[13] Maselli, G., & Nesticò, A. (2025). Machine learning algorithms and explainable artificial intelligence for property valuation. *Buildings*, *15*(15), 2678.

[14] Pita, R. P., de Carvalho, A. R., & Barbosa, R. M. (2026). A systematic review of the use of machine learning in the prediction of house pricing. *Journal of Housing Economics*, *62*, 101987.

[15] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144).

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

[17] Saiu, V., & Mocci, M. (2024). Explainable AI for urban real-estate prediction: A machine-learning framework for urban decision support. *Sustainability*, *16*(12), 5120.

[18] Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*(2), 2843-2852.

[19] Wang, Y., & Li, Z. (2025). Exploring housing price dynamics in sustainable cities through a cooperated big data driven machine learning method: Case study on a typical city in China. *Sustainable Cities and Society*, *110*, 105789.

[20] Ye, Y. (2022). An explainable model for the mass appraisal of residences: The application of tree-based machine learning algorithms and interpretation of value determinants. *Cities*, *128*, 103803.