Article

Comparative Study of Machine Learning Models for U.S. Housing Price Prediction

Wenguang Zhou [1],* and Wenjiao Zhou [2]

[1] Szkoła Główna Handlowa, Poland
[2] Akademia Leona Koźmińskiego, Poland
* Correspondence: Wenguang Zhou, Szkoła Główna Handlowa, Poland

Abstract: This study compares machine learning models used to predict US house prices using a large-scale real estate dataset covering all US states, cities, and zip code regions. The goal is to evaluate prediction accuracy and identify effective preprocessing strategies for high-cardinality location features. After cleaning the data through missing value handling, deduplication, and outlier removal based on interquartile range (IQR), leakage-aware feature engineering was applied, including aggregating zip codes into ZIP3, grouping by the top K cities, and date decomposition, while excluding target derived variables. This study trains and evaluates three models-linear regression, random forest, and XGBoost-on a reserved test set using mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2). The results show that XGBoost achieves the best performance, outperforming linear regression and random forest, and feature importance indicates that location indicators play a dominant role in prediction gain. The findings demonstrate that the XGBoost model outperforms linear regression and random forest models in predicting US house prices.

Keywords: housing price prediction, machine learning, regression models, feature engineering

Received: 31 December 2025
Revised: 23 February 2026
Accepted: 07 March 2026
Published: 14 March 2026

1. Introduction

House price forecasting is a core issue in household wealth assessment, mortgage approval, and investment decisions, and has gradually shifted from purely econometric or hedonic models to machine learning (ML) methods to better capture nonlinear patterns and complex feature interactions in real-world markets. Recent studies have consistently shown that forecasting performance is highly dependent on data preprocessing, feature engineering, and model selection. Previous research has demonstrated that systematic feature engineering can significantly improve the accuracy of house price forecasting and that ensemble learning methods often show strong predictive performance [1]. Another benchmark study reported that combining feature engineering with ensemble learning can effectively improve the prediction of house sale prices [2]. Further research indicated that advanced machine learning techniques, particularly boosting and hybrid or stacked methods, can outperform standard baseline models and improve forecasting accuracy [3]. Other studies focusing on AI-driven regression models compared multiple machine learning regressors and confirmed that modern nonlinear models can achieve high explanatory power in house price prediction [4]. From the perspective of regression techniques, prior work emphasized that a range of regression models, from linear to nonlinear, remain competitive when appropriate adjustments are applied [5]. Additional studies highlighted that identifying significant features is not only important for interpretability but is also directly related to improving prediction accuracy [6]. Evidence has also shown that support vector regression (SVR) can effectively model house prices when kernel functions and parameters are carefully selected [7]. Further findings suggest that machine learning approaches based on decision trees and boosting algorithms generally achieve higher prediction accuracy than simpler methods in practical applications [8]. Comparative evidence indicates that cross-model evaluation is essential because performance rankings may vary depending on dataset characteristics and evaluation metrics [9]. It has also been emphasized that establishing transparent baseline models, such as linear regression and decision tree models, is important before adopting more complex learners [10]. Based on these findings, this study uses real estate sales data from multiple regions to compare and evaluate representative machine learning models for house price prediction. The aim is to quantify the accuracy and robustness of different model families while ensuring that preprocessing and feature engineering procedures are consistent with commonly accepted practices in existing research.

2. Methodology

2.1. Dataset description

The empirical analysis is based on the USA Real Estate Dataset published on Kaggle, containing 2,226,382 U.S. real-estate listing records with property attributes and location fields (state/ZIP code) [11]. Each record includes property attributes and location information. In our setting, price is the dependent variable y, while the remaining fields are treated as predictors x. The main variables are presented in Table 1.

Table 1. Description of variables.

Feature Name	Type	Description
price	float64	current listing price (or recently sold price if sold recently)
status	object	listing status (e.g., ready for sale / ready to build)
bed	float64	number of bedrooms
bath	float64	number of bathrooms
acre_lot	float64	land area (acres)
house_size	float64	living area (square feet)
brokered_by	float64	encoded broker/agency identifier
street	float64	encoded street identifier
city	object	city name
state	object	state name
zip_code	float64	ZIP code
prev_sold_date	object	previous sold date (date/time)

To mitigate the impact of extreme values on house price forecasts, particularly for tree-based models and error-based indicators, outlier detection and removal were performed on numerical variables. First, box plots were used to visually examine distributions and potential outliers. Next, robust interquartile range (IQR) rules were applied to the main continuous variables, including price, number of beds, number of bathrooms, acreage, and house size. For each variable, the 25th percentile Q1, 75th percentile Q3, and IQR = Q3 - Q1 were calculated, and observations outside the interval [Q1 - 1.5 × IQR, Q3 + 1.5 × IQR] were removed. Since zip_code is a geographic identifier rather than a continuous measurement, the IQR rule was not applied to zip_code. After IQR-based filtering, the size of the dataset was further reduced from the imputed and deduplicated samples, ultimately retaining approximately 1,061,929 records (10 variables) for modeling, thereby improving training stability and generalization performance.

2.2. Data preprocessing

First, missing values were quantified using column counts and percentages to identify severely incomplete fields. The dataset had a large number of missing values (≈33%) in `prev_sold_date`, and significant missing rates were also observed in key numerical predictor variables such as `house_size`, `bath`, `bed`, and `acre_lot`. To maintain geographical consistency while preserving sample size, a stratified median imputation strategy was used for numerical variables. Specifically, the `bed`, `bath`, `house_size`, `acre_lot`, and `price` columns were imputed using the median within the same zip code region. For records lacking a zip code, a second stage of imputation was performed using the group median within the same state. Furthermore, the `brokered_by` column was removed from the modeling dataset because of its limited interpretability as a privacy-coded identifier and its relatively low benefit in handling missing values compared to core structural and location features. After this imputation, the remaining missing entries (primarily related to `prev_sold_date` and a small number of residual NAs) were removed using full case deletion to ensure the modeling tables were fully observable. This step generated a cleaned dataset containing 1,491,716 records with a 0% missing value among the retained variables, followed by deduplication and outlier filtering.

The cleaned data contains both numerical and categorical variables. Numerical variables include, for example, the number of beds, bathrooms, land area, and house size; categorical variables include street, city, state, and last sale date. For ease of modeling, categorical variables are one-hot encoded.

2.3. Machine learning models

Linear Regression: This study uses linear regression as an interpretable baseline model, assuming a linear relationship between house prices and input features. It estimates the coefficients of each predictor variable (e.g., number of bedrooms, number of bathrooms, house size, location-related variables) and predicts house prices as a weighted sum of these features.

Random Forest: Compared to a single decision tree, random forests reduce overfitting and enable them to capture non-linear relationships and feature interactions common in real estate pricing (e.g., the combined effect of location, house size, and lot size). Random forests also help assess feature importance, thus enhancing interpretability at the practical application level.

XGBoost: Through gradient-based optimization and regularization, XGBoost typically achieves high accuracy on structured/tabular datasets and is well-suited for large-scale housing data with mixed numerical and categorical features. In this study, XGBoost, as an advanced tree ensemble model, promises competitive predictive performance across US states.

2.4. Performance metrics

To evaluate the performance of the regression model in predicting US house prices, this study used three standard metrics: R^2, mean squared error (MSE), and mean absolute error (MAE). These metrics comprehensively reflect the goodness of fitness and the magnitude of prediction error from different perspectives.

Measures how much of the variance in the observed housing prices can be explained by the model. A higher R^2indicates better overall fit.

The coefficient of determination (R^2) can be expressed as:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Where:

$$SS_{\text{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 , \quad SS_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The Mean Squared Error (MSE) is computed as follows:

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

The Mean Absolute Error (MAE) can be written as:

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

Where:

- n denotes the total number of observations (properties) in the evaluation set.
- i indexes an individual observation, with $i = 1,2,\ldots,n$.
- y_i is the actual housing price of the i-th observation.
- $\hat{y}_i$ is the predicted housing price produced by the model for the i-th observation.
- $\bar{y}$ is the mean of the actual housing prices in the evaluation set, i.e., $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- SS_{res} is the residual sum of squares (sum of squared prediction errors).
- SS_{tot} is the total sum of squares (total variance of the target around its mean).

3. Empirical analysis

3.1. Exploratory Analysis of Features and Data Characteristics

After data cleaning, the dataset contains both numerical and categorical attributes. The main numerical variables include price, number of bedrooms, number of bathrooms, plot area, house size, and privacy-coded identifiers. Categorical variables include house status, city, state/province, and last sold date.

Figure 1 shows a Pearson correlation heatmap of the numerical subset. The heatmap shows a moderate positive correlation between price and the number of bathrooms (≈ 0.39) and house size (≈0.36), but a weaker correlation with the number of bedrooms (≈ 0.23), indicating that house structure characteristics are an important linear driver of house prices. Furthermore, strong correlations exist among the predictor variables, particularly between house area and number of bathrooms (≈0.65) and house area and number of bedrooms (≈0.60), suggesting potential multicollinearity that could affect the stability of coefficients in the linear model. In contrast, the correlation between street address and price is close to zero (≈-0.02), consistent with its characteristics as a coded identifier and a finite linear signal. Additionally, the derived feature price_per_sqft is highly correlated with price (≈ 0.78); However, since it is calculated using the target value (price/house area), it suffers from target value leakage and should not be used as an input feature for prediction.

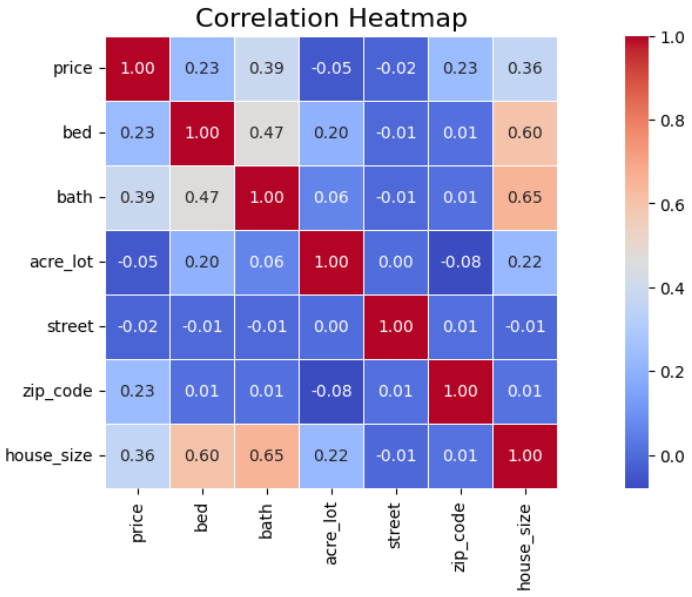

Figure 1. Correlation heatmap.

3.2. Model comparison

To ensure computational feasibility and reduce sparsity, high-cardinality location-related variables were compressed before one-hot encoding. All encoding rules were derived from training splits and applied to test splits to avoid information leakage. This enables machine learning models to be efficiently trained and effectively compared on the US housing dataset.

Table 2 compares the predictive performance of linear regression, random forest, and XGBoost on the preserved test set using MAE, MSE, and R² metrics.

Table 2. Model performance comparison.

Model	MAE	MSE	R^2
Linear Regression	93161.32	17327247942.10	0.668
XGBoost	79971.57	13490890828.81	0.741
Random Forest	128144.62	28433195120.71	0.455

Overall, XGBoost performs best across all three metrics, exhibiting the lowest error (MAE = 79,971.57, MSE = 13,490,890,828.82) and the highest goodness of fit (R² = 0.741). This indicates that the XGBoost ensemble model is more effective than other models in capturing nonlinear relationships and feature interactions in the housing market. This result is expected considering the significant geographical heterogeneity and structural interactions in the dataset (e.g., the joint effects of living space, number of bathrooms, and location proxy variables).

Linear regression provides a competitive baseline model (MAE = 93,161.32, MSE = 17,327,247,942.10, R² = 0.668), showing that most price fluctuations can be explained by the approximately linear effects of the constructed and encoded predictors. However, its performance consistently lags behind XGBoost, reflecting the limitations of the purely linear assumption under significant nonlinearity and interactions (especially between classification encodings related to house size, number of bedrooms, number of bathrooms, and location).

In contrast, random forest performs the worst among the three models (MAE = 128,144.62, MSE = 28,433,195,120.71, R² = 0.455). One plausible explanation is that random forests, as a bagging-based ensemble model, may require more and/or deeper trees to achieve performance comparable to gradient boosting on large-scale heterogeneous tabular data, and their effectiveness decreases when the signal is primarily influenced by complex location effects and sparse class encoding. Furthermore, the configuration of this model (and the training subsampling, possibly for computational feasibility) further limits its ability to fit fine-grained patterns compared to gradient boosting.

The findings of this study are consistent with previous comparative studies. For example, previous research has generally shown that boosting-based tree ensemble methods, such as XGBoost or LightGBM, often outperform random forests and linear baseline models in house price prediction because they are able to model nonlinear patterns and complex feature interactions [2-4,8].

3.3. Feature importance analysis

Figure 2 illustrates the most important features extracted from the XGBoost model using the gain criterion. The results show that location-related variables dominate the importance ranking, with most top-ranking features corresponding to one-hot encoded indicators of state, city, and ZIP3. Most notably, California has the largest gain, with a significant advantage compared to other states, indicating that California housing information can significantly improve prediction accuracy. Indicators from several other states, such as Ohio, Massachusetts, Washington, and Oklahoma, as well as more granular location proxies, ZIP3 zip codes, and specific cities like Seattle, also appear among the most influential features. This pattern suggests that geographic heterogeneity is a major driver of price differences in multi-state datasets and highlights the importance of effectively representing location effects. Although structural variables such as living space and number of rooms are known predictors, their absence from the gain-based importance ranking does not imply their insignificance; rather, it suggests that the model first segments the data along geographic boundaries and then refines the data within regions using structural attributes, resulting in greater marginal improvements. Overall, the feature importance analysis further confirms that accurate prediction of U.S. house prices requires models that can capture complex location-related patterns, consistent with the superior performance of XGBoost observed in the model comparison.

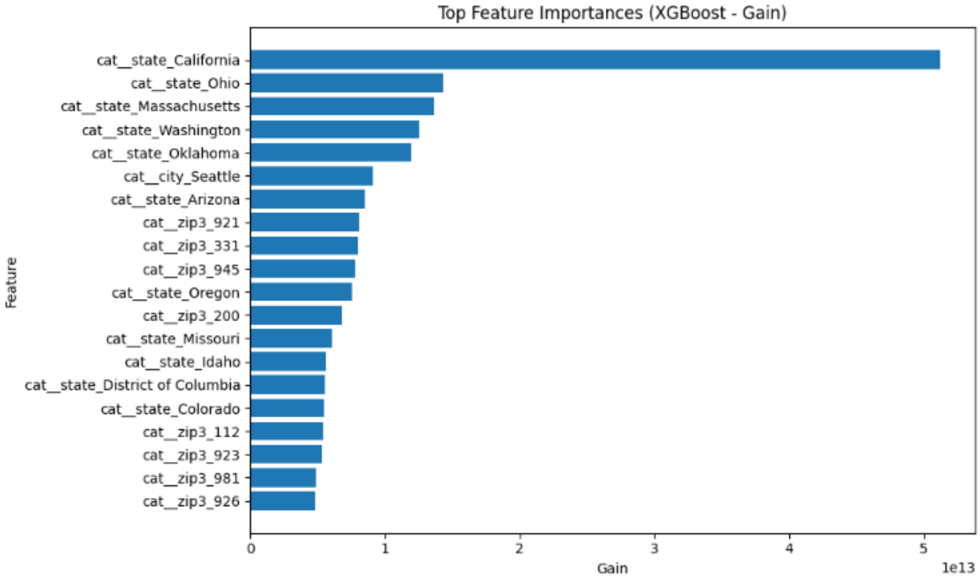

Figure 2. Top 20 feature importances of the XGBoost model measured by gain.

4. Discussion

Based on the test set results in Table 2, the XGBoost model performed best (MAE = 79,971.57, MSE = 13,490,890,828.82, $R^2 = 0.741$), outperforming linear regression ($R^2 = 0.668$) and random forest ($R^2 = 0.455$). This advantage is mainly attributed to XGBoost's gradient boosting framework, which corrects previous errors by sequentially building decision trees and combines regularization to improve generalization ability. In a multi-state housing dataset in the United States with significant geographical heterogeneity and nonlinear interactions-for example, the joint effects of living space, number of bathrooms, and location proxy variables-ensemble models based on boosting algorithms are well-suited to capturing complex patterns that linear models cannot represent, while models based on bagging algorithms (random forests) may not be able to learn these patterns efficiently. The poor performance of random forests is also consistent with the fact that when dealing with large-scale heterogeneous tabular data, random forests usually require larger model capacity and more fine-tuning to match boosting algorithms, and they may be more sensitive to sparse or high cardinality class representations.

Feature engineering is effective because it directly addresses two key challenges of this dataset: high cardinality of categorical variables and information leakage. First, location-related attributes such as postal codes, cities, and states contain the strongest predictive signals, but simply one-hot encoding the original postal codes or dates results in a highly dimensional sparse matrix, which is computationally inefficient and prone to overfitting. By aggregating postal codes into ZIP3 and classifying rare city categories into the "Other" category (Top K strategy), the feature space is reduced to a manageable size. For example, the processing matrix for the training set is (858,653, 1,450), and the processing matrix for the test set is (212,664, 1,450), thus achieving efficient model training while preserving most of the geographical signals. Second, converting `prev_sold_date` into an interpretable time component avoids the explosive growth at the category level and captures temporal information in a structured way. Finally, removing the price per square foot from the model features prevents target value leakage, as it is calculated based on the target price and house area; failing to remove it would artificially inflate model performance. Gain-based feature importance further validates the feature engineering strategy: the most influential predictors are primarily state, city, and zip code (ZIP3) indicators, confirming that engineered location representation can capture key price drivers.

The importance of model ranking and the observed location- and size-related predictors is highly consistent with previous comparative studies. In particular, previous studies have shown that ensemble models based on boosting algorithms (such as XGBoost) generally outperform random forests and linear baseline models in house price prediction tasks because they are more effective at capturing nonlinear relationships and feature interactions [2-4,8]. In addition, prior research has emphasized that predictive performance largely depends on feature engineering, and structural and location-related variables are among the most important drivers [1,6]. Therefore, the results of this study are consistent with the broader literature: well-designed representations of location and property structure, combined with gradient boosting models, can provide strong predictive performance for large-scale house price datasets.

5. Conclusion

In summary, with the rapid increase in data availability and the widespread application of machine learning in real estate analytics, house price forecasting increasingly relies on advanced algorithms and scalable feature engineering techniques. This study uses large-scale real estate sales or listing data from multiple US states to compare three representative models: linear regression, random forest, and XGBoost. This study establishes an easily tractable modeling framework and achieves robust predictive performance. Empirical results show that XGBoost achieves the highest overall accuracy, outperforming linear baseline models and bagging-based ensemble models. Further, important analysis further confirms that geographic indicators (state/city/ZIP3) are the main driver of improved predictive performance in heterogeneous national markets.

From a practical perspective, the results demonstrate that categorical, scalable feature engineering combined with gradient boosting models provides a deployable solution for real-world valuation and risk assessment tasks. For practitioners such as real estate platforms, financial institutions, and policy analysts, the proposed process offers an efficient method to generate cross-state consistent price estimates, support pricing strategies, improve loan and collateral risk assessment, and enhance market monitoring. In the future, integrating richer variables related to property quality and community levels, such as year of construction, renovation status, school quality, and local economic indicators, and employing geographically aware validation strategies can further enhance the model's generalization ability and decision relevance. Only by combining robust data governance, assessment of perceived leaks, and scalable modeling techniques can house price prediction systems better support transparent, reliable, and sustainable decision-making in the evolving digital real estate economy.

References

1. D. J. C. Sihombing, "Application of Feature Engineering Techniques and Machine Learning Algorithms for Property Price Prediction," *JITSI: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 5, no. 2, pp. 72-76, 2024. doi: 10.62527/jitsi.5.2.241
2. O. E. Ogunbiyi, "House Sale Price Prediction using Feature Engineering Techniques and Ensemble Learning Algorithms (Doctoral dissertation, Dublin, National College of Ireland)," 2020.
3. Q. Truong, M. Nguyen, H. Dang, and B. Mei, "Housing price prediction via improved machine learning techniques," *Procedia Computer Science*, vol. 174, pp. 433-442, 2020. doi: 10.1016/j.procs.2020.06.111

4. M. S. V. Tyagadurgam, V. N. Gangineni, S. Pabbineedi, A. B. Kakani, S. K. K. Nandiraju, and S. K. Chundru, "Using Artificial Intelligence-Based Machine Learning Regression Models for Predictions of Home Prices," *European Journal of Applied Science, Engineering and Technology*, vol. 3, no. 3, pp. 404-416, 2025. doi: 10.59324/ejaset.2025.3(3).29

5. J. Manasa, R. Gupta, and N. S. Narahari, "Machine learning based predicting house prices using regression techniques," In *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*, March, 2020, pp. 624-630. doi: 10.1109/icimia48430.2020.9074952

6. M. R. Saefudin, M. R. Putri, A. Hadi, H. Wijayanto, and B. Irmawati, "Significant Features for House Price Prediction Using Machine Learning," In *2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, November, 2024, pp. 659-664. doi: 10.1109/comnetsat63286.2024.10862860

7. J. Y. Wu, "Housing price prediction using support vector regression," 2017. doi: 10.31979/etd.vpub-6bgs

8. R. Monika, J. Nithyasree, V. Valarmathi, G. R. Hemalakshmi, and N. B. Prakash, "House price forecasting using machine learning methods," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 11, pp. 3624-3632, 2021.

9. I. C. Obagbuwa, and S. Danster, "Housing Price Prediction Using Machine Learning Techniques," In *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*, April, 2024, pp. 1-12. doi: 10.1109/seb4sdg60871.2024.10629723

10. M. Thamarai, and S. P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning," *International Journal of Information Engineering & Electronic Business*, vol. 12, no. 2, 2020. doi: 10.5815/ijieeb.2020.02.03

11. G. Cabrera, J. D. Díaz, and E. Hansen, "Real Estate Returns and the Macroeconomy: Insights from Big Data in the US, Canada, and the UK," *The Journal of Real Estate Finance and Economics*, pp. 1-89, 2026.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of GBP and/or the editor(s). GBP and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.