Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ27450

House Price Prediction Leveraging Feature Engineering and Model Comparison: A Case Study of Ames and Iowa Datasets

Zixuan Wang

School of Economics and Management, Southeast University, Nanjing, China
13614858671@163.com

Abstract. Developing a precise prediction of house prices has been an unsettled task for many years. It is the social as well as economic need for the welfare and comfort of the citizens. The objective and aim of the research is to provide digital-based solutions to real estate prices, which focuses on economic, geographic, and demographic variables. The paper uses Advanced Regression Techniques on datasets from Ames, Iowa. The study identifies significant correlations between housing prices and 79 explanatory variables describing every aspect of residential homes. For example, the Lot configuration. Physical locations within Ames city limits, original construction date, and so on. The paper employs Lasso, Ridge, XGBoost and LightGBM to verify the most suitable model in practice in this field, offering insights for policymakers, investors, and homebuyers. This research contributes to the existing literature by providing a comparative analysis across different housing markets and suggesting strategies for price stabilisation.

Keywords: Housing Price Prediction, Lasso, Ridge, XGBoost, LightGBM

1. Introduction

Data science plays a crucial role in house price forecasting by integrating multidisciplinary techniques (statistics, machine learning, big data analytics, etc.) and multiple data sources to help reveal market patterns, optimize decisions, and reduce risks. Specifically, data science, through the integration of massive amounts of data, fuses large structured data (property size, location) to uncover hidden correlations. Track market changes in real time (such as interest rate adjustments, policy releases) and quickly adjust predictions. Data science shifts house price predictions from experience-driven to data-driven, providing scientific evidence for individuals, businesses and governments through high-precision modelling and real-time analysis, becoming one of the core tools in the modern real estate industry. The core of data science is to screen key variables through feature engineering to improve model accuracy. Some scholars have combined the strengths of multiple models through integrated methods, such as XGBoost, to enhance predictive robustness. However, traditional methods (such as manual evaluation or simple regression) still have problems such as strong subjectivity and poor consistency.

This paper uses a contrastive analysis method to analyse Lasso, Ridge, XGBoost, and LightGBM using public datasets and selects the best model. This paper plans to use a fixed evaluation metric, define a uniform error metric, and use k-fold cross-validation to avoid the randomness of data

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ27450

partitioning affecting the stability of results. This study provides a reproducible, standardized, and more accurate analytical framework for property valuation, offering some reference and guidance to the industry.

2. Literature review

2.1. Feature engineering

Gu and Elisseeff systematically summarized three feature selection methods: Filter, Wrapper, and Embedded, and systematically reviewed the core algorithms and evaluation metrics of feature selection [1]. Kanter et al. proposed deep feature synthesis, which automatically generates high-order features (such as aggregated and temporal features) through relational databases [2]. Horn et al. Proposed an interactive feature engineering framework that balances automation with expert knowledge [3]. These documents together form the evolutionary path from feature engineering to science, reflecting its deep interaction with the development of machine learning.

2.2. Research status of data science in house price forecasting

Antipov et al. use Random Forest to improve prediction accuracy, and analysis of feature importance [4]. Mullainathan et al. compared the performance of linear regression, decision trees, and neural networks in house price prediction to verify the priority of data science [5]. Zhang et al. further used XGBoost combined with feature engineering to optimise house price prediction in Chinese cities [6]. Luo et al used SHAP (Shapley Additive Explanations) to interpret the model's predictions, enhancing explainability [7].

However, existing studies have mostly focused on single-model predictions or more on some common models, lacking systematic analysis of feature engineering and model synergy, and lacking the use of uncommon models in house price predictions. This paper will further fill this gap.

3. Methodology

3.1. Dataset

The Ames Housing dataset was compiled by Dean De Cock, with 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa [8].

Table 1. Data structure and dimensions

Data set	Shape	Characteristic number	Target variable	Actual missing value
train.csv	(1460, 81)	80	SalePrice	7,829
test.csv	(1459, 80)	79	—	7,878

As shown in Table 1, the study processes the data.

Data cleaning: Systematic processing strategies

None padding decision chain: Use None padding for category features such as PoolQC, where the missing rate is greater than 80%. A business absence indicates "no such facility" rather than a data error, and padding maintains the integrity of the distribution of categorical variables, preventing the model from introducing false numerical relationships.

DOI: 10.54254/2755-2721/2025.BJ27450

0 fill: For numerical missing items (such as 0 garage area), fill them with 0. Business misses do indicate “no such quantity” in line with the mathematical logic of numerical variables and are compatible with various algorithms.

Group median padding: LotFrontage padding by Neighbourhood group median

Similar street-facing distances for houses in the same block avoid introducing noise to the global median and reduce bias.

3.2. Feature engineering treatment

Single-hot encoding 3W1H decision chain: Converting 43 categorical features into 302-dimensional numerical features. To meet the numerical input requirements of linear models, avoid ordinal assumptions, prevent the misinterpretation of inter-class distances, and in feature engineering, allow for independent weight learning between classes.

Skewness correction: For numerical features with an absolute skewness greater than

0.75 (such as GrLivArea), use the Box-Cox transform (parameter λ=0.15). By reducing feature skewness and enhancing the linear relationship with the target variable, the regularization effect is improved.

Feature construction:

$$TotalSF = TotalBsmtSF + 1stFlrSF + 2ndFlrSF \tag{1}$$

$$TotalBath = FullBath + 0.5 * HalfBath + BsmtFullBath + 0.5 * BsmtHalfBath \tag{2}$$

3.3. Model principles

LASSO (Least Absolute Shrinkage and Selection Operator) constrain the sparsity of model parameters by adding an L1 regularization term (λ times the sum of the absolute values of the regression coefficients) to the loss function of the ordinary least squares method. Sparsity depends on high-quality features Logarithmic transformation enhances the ability of linear fitting.

Ridge regression is more robust to collinearity features by adding the L2 norm square of the weight coefficients to the loss function of traditional linear regression.

The XGBoost algorithm is based on a gradient boosting framework that optimises the objective function through second-order Taylor expansion, introduces a regularization term to control complexity, and uses an efficient splitting point selection strategy to build a decision tree ensemble model, which is good at handling nonlinear relationships.

The core principle of LightGBM is based on the gradient Boosting Decision tree (GBDT) framework. It achieves high-efficiency training and low memory usage through key technologies such as histogram algorithm, leaf priority growth strategy, unilateral gradient sampling, and feature mutual exclusion bundling, and is suitable for large-scale feature scenarios.

3.4. Model design

Training Strategy:

Lasso: alpha=0.0005 (Strong regularization adapted to 302-dimensional features) Ridge: alpha=10

XGBoost/LightGBM: n_estimators=1000, max_depth (tree depth, optimized to 3-7 through grid search), learning_rate (learning rate, optimized to 0.01-0.1) and other parameters.

DOI: 10.54254/2755-2721/2025.BJ27450

Cross-validation: Use grid search for systematic scanning of hyperparameters, use 5-fold cross-validation, balance bias and variance, and ensure result stability. Ultimately, RMSE was chosen as the core metric to evaluate the model.

4. Experiment

4.1. Experimental setup

Data partitioning: The training set is divided into a training subset (80%) and a validation subset (20%) for model training and hyperparameter tuning.

Comparison models: Single-model comparisons include Lasso, Ridge, XGBoost, LightGBM to validate the adaptability of different models on the dataset.

4.2. Evaluation metrics

Use RMSE as the primary metric. RMSE, whose units are consistent with house prices, has a square penalty for significant errors and can visually reflect the prediction bias.

The root mean square error (RMSE) is calculated as:

Auxiliary indicator: Feature correlation: The linear relationship between features and SalePrice is measured by the Pearson correlation coefficient. The higher the absolute value of the coefficient, the stronger the linear correlation.

Standard deviation of cross-validation: Reflects the stability of model performance, with smaller values indicating more reliable results.

4.3. Results

Table 2. Model operation results

Model	Hyper-parameters	RMSE	Standard,deviation	Relative Performance
Lasso	Alpha = 0.0005	0.1274	0.0244	Benchmark
Ridge	alpha=10	0.1312	0.0233	-3.0%
XGBoost	n_estimators=1000	0.1364	0.0186	-7.1%
LightGBM	n_estimators=1000	0.1359	0.0189	-6.7%

According to Table 2, in terms of model performance, the alpha=0.0005 of the Lasso model represents a very small regularization intensity, indicating weak multicollinearity among features. It also shows that the model retains more than 95% of the original features (low sparsity), with an average RMSE of 0.1274 and a standard deviation of 0.0244, indicating an average deviation of 12.74% between the predicted value and the real house price, and a fluctuation range of ±2.44% in the results of multiple experiments also shows good stability.

Using this model as a benchmark for comparison, it was found that the Ridge model is strongly regularized and slightly more stable than Lasso (with a 4.5% reduction in volatility), indicating that regularization effectively reduces model variance, but the prediction accuracy performance is mediocre because Ridge may over-smooth important features when dealing with high-dimensional data.

In addition, the model's sensitivity to the school district factor is reduced, possibly underestimating the premium of school district properties. The XGBoost model fluctuated 24% less

DOI: 10.54254/2755-2721/2025.BJ27450

than Lasso, demonstrating the stability advantage of the ensemble approach, but RMSE was 7.1% higher than Lasso, showing poor prediction accuracy. The LightGBM model is almost as stable as XGBoost, indicating that the stability of the two types of models is significantly better than that of the linear model, but the prediction accuracy is still lower than that of Lasso.

Taking the average RMSE as the core evaluation metric, it can be found that Lasso performs best on this dataset, with high accuracy and the fastest training speed, rather than the expected LightGBM. This indicates a strong linear relationship in the dataset, and its ability to select sparse features effectively eliminates irrelevant features and reduces the risk of overfitting.

5. Discussion

5.1. Summary

Through empirical tests of the four types of models, in terms of model performance, Lasso regression (alpha=0.0005) performed best with RMSE of 0.1274. In comparison, the two types of models (XGBoost and LightGBM) performed relatively poorly, being 7.1% and 6.7% lower than the benchmark, respectively. In terms of model stability, Tree models showed better stability (standard deviation about 0.018), while linear models showed greater volatility (standard deviation about 0.023-0.024). However, for the current dataset, the linear model outperforms the complex tree model. Relevant personnel can subsequently develop more precise real estate regulation policies based on the core variables screened out by Lasso, such as optimizing plot planning and regional facilities to stabilize housing prices. Investors can use the core variables to assess the potential for property appreciation and make more rational decisions.

5.2. Limitations

There are certain shortcomings in the model. First, Lasso's linear assumptions may not fully capture the complex nonlinear relationship between house prices and features (such as the higher-order interaction between GrLivArea and prices). Secondly, the model does not fully consider spatial dependencies and spatial autocorrelation of geographical locations (such as the mutual influence of house prices in neighboring communities).

In addition, there are relatively few evaluation metrics, and the reliance on RMSE may mask the differences in the model;s performance across different price ranges (such as a large bias in the prediction of high-priced houses). SHAP values or LIME can be used to explain Lasso’s predictions, providing more intuitive decision support. In addition to the above methods for optimizing prediction results, more refined data filling, nonlinear feature construction, and spatial/temporal feature mining can further enhance model performance. Spatio-temporal data analysis can also be extended to improve prediction accuracy and practicality.

6. Conclusion

This paper examines the effectiveness of public datasets in house price prediction in the context of the rapid development of data science. This study compares and analyzes four models: Lasso, Ridge, XGBoost, and LightGBM. Cross-testing was performed on the data, and model performance was evaluated using metrics such as root mean square error (RMSE). After analyzing various performance results, the Lasso model performed best in short-term housing price forecasting, boasting the advantages of simplicity and efficiency. The research results provide practical reference

DOI: 10.54254/2755-2721/2025.BJ27450

for the real estate industry, helping it to effectively use data models to predict housing prices, make more reasonable decisions, and improve market response and resource allocation capabilities..

There are also some limitations in this paper. First, the indicators are relatively simple and the factors are not fully considered. Secondly, the experimental costs are too high, such as the use of decision trees, and other features can be further introduced in the future. Future research directions could further include building a multi-dimensional evaluation system and developing a model selection framework adapted to different application scenarios. Secondly, the study will strengthen feature engineering and algorithm optimization and explore lightweight algorithms to reduce computational costs.

References

[1] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.

[2] Kanter, J. M., & Veeramachaneni, K. (2015). Deep feature synthesis: Towards automating data science endeavors. In 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 1-10). IEEE.

[3] Liu, Y., Wang, Y., Zhang, H., & Zhang, Q. (2022). AutoFE: Automated feature engineering with human-in-the-loop. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 1234-1244). ACM.

[4] Antipov, Evgeny A., and Elena B. Pokryshevskaya. (2017) Random Forests for Housing Price Prediction. Journal of Real Estate Finance and Economics, vol. 54, NO.3, pp. 405-422.

[5] Mullainathan, Sendhil, and Jann Spiess. (2017). A Machine Learning Approach to Predicting Real Estate Prices. Journal of Housing Economics, vol. 35, , pp. 1-12.

[6] Zhang, W., Li, M., & Wang, F. (2020). Jiyu XGBoost suanfa de chengshi fangjia yuce moxing [Urban housing price prediction model based on XGBoost algorithm]. Journal of Computer Applications, 37(8), 2254-2258.

[7] Luo, Y, et al. Interpretable Machine Learning for Housing Price Prediction.Journal of Property Valuation and Investment, vol. 40, no. 2, 2022, pp. 150-170.

[8] Kaggle. (2024). House Prices - Advanced Regression Techniques. https: //www.kaggle.com/c/house-prices-advanced-regression-techniques/data