Bayesian Optimization of Lasso and XGBoost Models for Comparative Analysis in Housing Price Prediction

Runze Zheng*

Pamplin College of Business, Virginia Polytechnic Institute and State University, 24060, Blacksburg, the United States

Abstract. Fluctuations in housing prices have a profound impact on the broader economy and people's livelihoods. Accurate housing price predictions contribute to enhanced market transparency and the formulation of evidence-based policies. This paper focuses on optimizing two machine learning models, Lasso Regression and XGBoost, using Bayesian optimization for predicting housing prices. By leveraging economic features such as Average Earnings, Gross Domestic Product (GDP), Mortgage rates, Population, and Unemployment Rate, the models aim to improve prediction accuracy in the housing market. The Lasso model, known for its feature selection capability through L1 regularization, was fine-tuned using Bayesian optimization to minimize mean squared error (MSE). The XGBoost model, designed for handling large-scale, non-linear datasets, was also optimized using the same method. After optimization, the Lasso model achieved an MSE of 240,498,369.05 and an R^2 score of 0.977, while the XGBoost model showed superior performance with an MSE of 80,273,332.19 and an R^2 score of 0.9914. SHAP analysis was used to interpret the models, revealing that Average Earnings and GDP were the most influential features in both models. The results demonstrate that while both models perform well, XGBoost's ability to handle non-linearity and high-dimensional data makes it more effective in housing price predictions.

1 Introduction

In recent years, frequent fluctuations in real estate market prices have become one of the focuses of attention from all sectors of society. With the accelerating global urbanization process, population growth, and economic restructuring, changes in real estate prices not only profoundly affect the macro-control of the national economy and financial stability, but are also closely linked to people's livelihood, urban development, and investment decisions. Therefore, accurately predicting housing price trends is crucial for formulating policies, conducting economic regulation, and preventing market risks. Especially in the context of increasing global economic uncertainty and continuous macroeconomic policy adjustments, accurate housing price forecasting can help improve market transparency, reduce investment

* Corresponding author: runze@vt.edu

risks, and provide a scientific basis for the government to formulate more targeted policies [1].

In the field of housing price forecasting, scholars have explored a variety of methods and techniques. Early studies mostly used traditional statistical models based on economic theories, such as multiple linear regression (MLR), agent-based models (ABM), time series models (such as AutoRegressive Integrated Moving Average（ARIMA）), etc., which performed well with small data sets and relatively simple economic environments [2]. However, as the amount of data and the complexity of variables increase, traditional methods have difficulty capturing the complex nonlinearities and diversity in housing price data. Therefore, in recent years, the application of machine learning methods in housing price forecasting has gradually increased. Machine learning models such as Lasso and XGBoost models have achieved significant improvements in prediction accuracy compared to traditional statistical methods. Lasso is a linear model based on L1 regularization that can effectively perform feature selection to reduce the problem of multiple collinearity. XGBoost, as an ensemble model based on gradient boosting, has strong nonlinear modeling capabilities and efficient training performance. In addition, deep learning models (such as long short-term memory networks (LSTM) and convolutional neural networks (CNN)) have also shown strong capabilities in extracting nonlinear features from large-scale data. However, model optimization and applicability remain important challenges in current research due to the complexity of different types of data and feature selection. Therefore, how to select an appropriate prediction model and combine it with an effective parameter-tuning method remains a research focus in academia and industry practice.

This study aims to use Bayesian optimization techniques to tune and compare two different types of models—Lasso and XGBoost models—to improve the accuracy of housing price predictions. This paper first introduces the dataset used and its characteristics, performs data cleaning and preprocessing, and uses visualization methods to reveal the distribution of variables and their relationships. Next, the principles and optimization methods of the Lasso and XGBoost models are described in detail. Subsequently, the prediction performance of the models is evaluated experimentally and the advantages and disadvantages of the two models are compared and analyzed using relevant indicators. Finally, this paper discusses the limitations of the research and future research directions. Through comparative analysis, this study hopes to provide an efficient model selection and optimization strategy for real estate market price forecasting.

2 Data and methods

2.1 Data sources

The dataset utilized in this study was sourced from Kaggle and encompasses several key ec onomic and real estate-related variables, such as GDP, average income, and mortgage intere st rates (https://www.kaggle.com/datasets/theelahi/us-home-price-prediction?resource=dow nload). These variables include Average_Earning, GDP, Mortgage, Population, Unemploym ent_Rate, and the target variable Average_Sales_Price. These variables are widely recognize d in the literature as key determinants of housing prices, with GDP and income level being s trongly correlated with real estate market fluctuations [3].

During preprocessing, irrelevant fields such as 'DATE' were removed. Then all the feature variables were standardized to ensure that variables of different scales would not cause bias during model training. The standardization method used makes the mean of the features 0 and the standard deviation 1, eliminating scale differences between features. Outliers were handled using the Z-score method, where values exceeding three standard deviations were

eliminated to ensure data stability. For a small number of missing values, the mean-filling method was used to reduce the impact of missing values on the model.

Fig.1 shows the distribution of the target variable Average_Sales_Price. The housing price data show a clear right-skewed distribution, indicating that most housing prices are concentrated below $300,000, while prices for high-end properties are scarce but greatly affect the overall mean. Approximately 75% of properties are priced below $300,000, whereas the highest price reaches nearly $1 million. This distribution reflects the current state of the US housing market, where some high-end properties are worth much more than those of middle-income families [4].

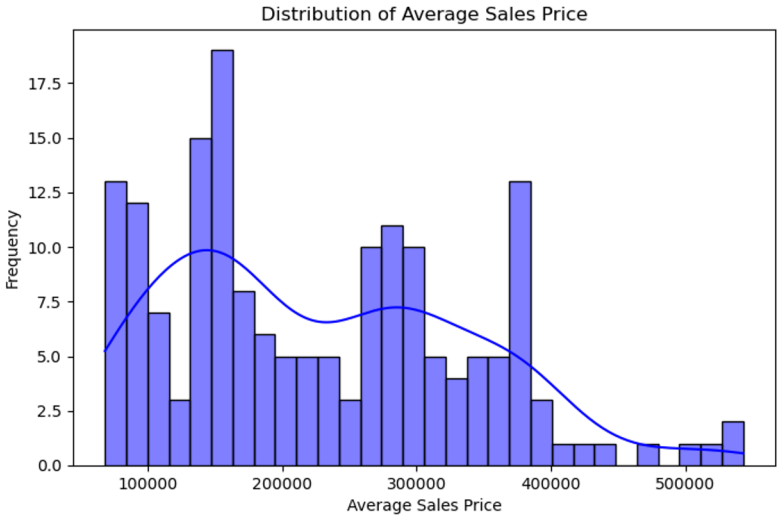

Fig.1. Sales_price_distribution (Photo/Picture credit: Original)

2.2 Methods and Models

This study focuses on optimizing two models, Lasso regression and XGBoost, through Bayesian optimization to enhance the accuracy of housing price predictions. Lasso regression employs L1 regularization to prioritize key features by eliminating those with low correlation, thus simplifying the model and mitigating overfitting. Lasso is particularly effective in handling high-dimensional data, improving model stability through its robust feature selection process.

XGBoost is a highly efficient, open-source machine learning framework, recognized for being over ten times faster than many mainstream solutions, particularly in distributed or memory-constrained environments. Its efficiency stems from innovations such as tree learning algorithms optimized for sparse data and the weighted quantile sketch algorithm, which efficiently handles the approximation of tree-based learning on sample weights. In addition, XGBoost accelerates model training through parallel and distributed computing, enabling efficient hyperparameter tuning for large datasets [3,5].

Bayesian optimization is leveraged, as it is particularly effective for models requiring rapid convergence, such as machine learning algorithms [3].

2.3 Evaluation Metrics

To comprehensively assess model performance, this paper utilizes three primary evaluation metrics: mean square error (MSE), coefficient of determination (R^2), and cross-validation

error. MSE evaluates the average squared difference between predicted and actual values, where lower values indicate higher prediction accuracy. R^2 reflects the proportion of variance in the data that the model explains, with values closer to 1 indicating stronger interpretive power [4]. Additionally, cross-validation error is employed to evaluate model stability through 10-fold cross-validation, ensuring consistency and robustness.

In addition, the SHAP tool is used to explain the importance of each feature in the model's prediction, providing deeper insights into the decision-making process of both Lasso and XGBoost models.

3 Results analysis

3.1 Data visualization analysis

To gain a deeper understanding of the relationship between key features and housing prices, the paper generated scatter plots and correlation heatmaps to visualize the interactions among variables and their influence on property prices. Fig.2 illustrates a strong positive correlation between GDP and housing prices, underscoring the significant influence of macroeconomic growth on property market dynamics [5]. This relationship highlights the sensitivity of housing prices to broader economic performance. This correlation is also verified in other economic characteristics, such as the strong correlation between population and house prices. The housing price distribution reveals that high-end property prices significantly exceed those of average residential properties, confirming the skewed distribution of the real estate market, where luxury properties disproportionately elevate the overall price levels. Fig.3 shows a correlation heat map, which shows that the correlation between population and GDP and Average_Sales_Price is high, indicating that income levels and macroeconomic conditions have an important impact on housing prices [6].

Fig.2. Gdp_vs_sales_price (Photo/Picture credit: Original)

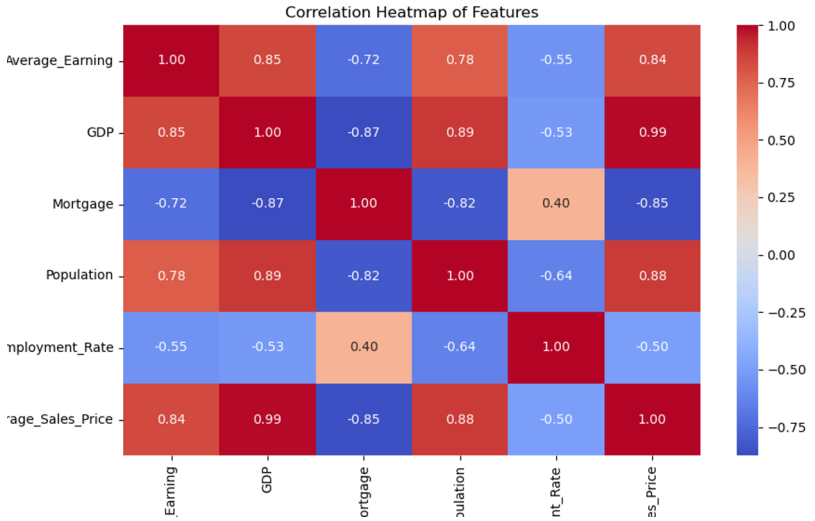

Fig.3. Correlation_heatmap (Photo/Picture credit: Original)

3.2 Forecast analysis and model performance evaluation

To comprehensively evaluate the predictive performance of the Lasso and XGBoost models in forecasting housing prices, this study performed a multi-dimensional assessment focusing on key evaluation metrics such as R^2, MSE, and the impact of feature selection. The following is a detailed analysis of the prediction results and model performance of the two models.

After Bayesian optimization, the Lasso model achieved an R^2 of 0.9779 and an MSE of 240,498,369.05, compared to a pre-optimization MSE of 242,635,261.31. These results indicate a significant reduction in prediction error following the optimization process. Lasso regression, as a linear model with L1 regularization, effectively reduces overfitting by eliminating less important variables, thus retaining only the most significant features relevant to housing price predictions. Therefore, the Lasso model excels at dealing with linear features. In this study, population and GDP were identified as key factors influencing housing prices. The correlation analysis revealed a strong positive relationship, with population exhibiting a correlation coefficient of 0.88 and GDP showing an even stronger correlation of 0.99 with housing prices (Fig.3). These findings are consistent with existing literature, which also highlights the significant impact of household income levels and macroeconomic growth on housing prices.

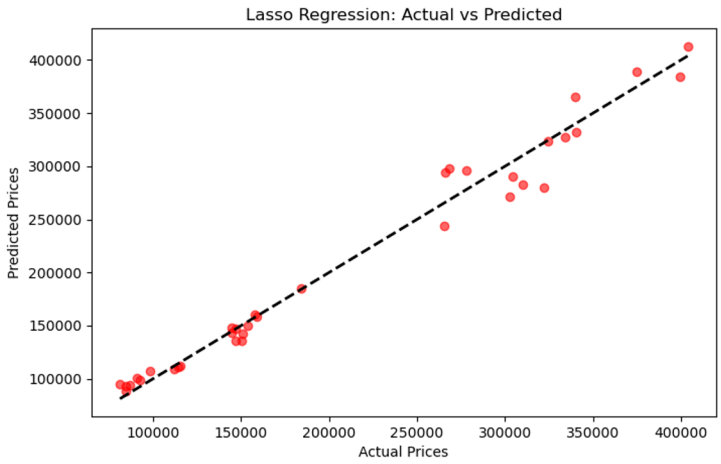

Fig.4. Lasso_actual_vs_predicted (Photo/Picture credit: Original)

Fig.4 presents the comparison between the predicted and actual housing prices using the Lasso model. The results indicate that the model performs more accurately in predicting medium- and low-priced properties, as the deviation between the predicted and actual values remains minimal. This finding suggests that the Lasso model exhibits high accuracy in explaining linear variations in house prices, particularly within the medium- and low-price ranges, enabling the model to capture overall pricing trends more effectively. Nevertheless, the Lasso model demonstrates limitations in predicting high-end properties. Owing to its linear structure, it struggles to capture intricate nonlinear relationships, resulting in relatively large prediction errors for high-priced properties. Conversely, XGBoost, an ensemble decision tree model based on gradient boosting, excels in capturing complex nonlinear relationships and demonstrates superior fitting capabilities [7]. In this research, the optimal hyperparameter configuration for XGBoost was obtained through Bayesian optimization. The optimized XGBoost model achieved an MSE of 80,260,789.90, compared to 97,009,990 before optimization, illustrating that XGBoost outperforms in handling high-dimensional and nonlinear datasets.

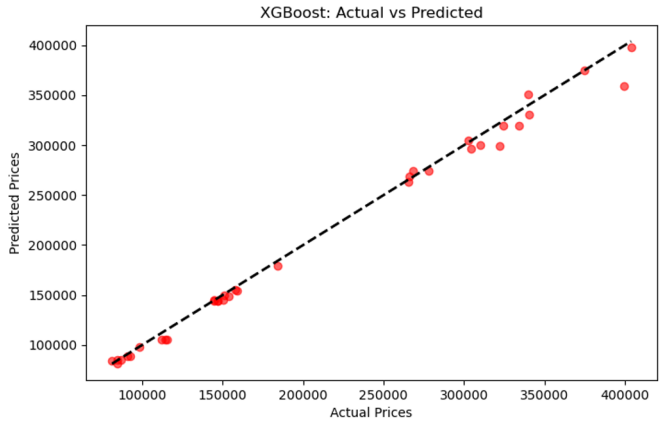

Fig.5. Xgb_actual_vs_predicted (Photo/Picture credit: Original)

Fig.5 presents the comparison between the predicted housing prices from the XGBoost

model and the actual housing prices. It is evident that the XGBoost model successfully captures complex feature relationships and provides an accurate fit, particularly for high-priced properties. Ensemble learning models like XGBoost offer distinct advantages over linear regression methods, particularly in handling complex non-linear relationships [8].

When predicting high-priced properties, the XGBoost model exhibited stable and precise performance, with minimal deviation between the predicted and actual values. This highlights its effectiveness in managing non-linear features. In comparison, the XGBoost model demonstrated a higher R^2 value of 0.9915 than the Lasso model, underscoring its superior capacity to capture housing price trends, thereby indicating a stronger fit and robustness to the data [9].

Furthermore, the stability of the model was evaluated through 10-fold cross-validation, where the cross-validation error of the XGBoost model was significantly lower than that of the Lasso model. This suggests better consistency and robustness. Across multiple cross-validations, the MSE of the XGBoost model stabilized at approximately 80,260,789.90, demonstrating that the model maintains high prediction accuracy and consistency across different data partitions. In contrast, the cross-validation MSE for the Lasso model exhibited greater fluctuations, averaging 240,498,369.05, indicating lower stability when faced with different datasets.

To further explore the contribution of each feature to the model's prediction, the SHAP (Shapley Additive Explanations tool was employed to interpret the model. SHAP values offer a clear explanation of the prediction process and assist in identifying which features significantly impact the final prediction. Fig.6 and Fig.7 display the SHAP value distributions of the most important features in the Lasso and XGBoost models, respectively.

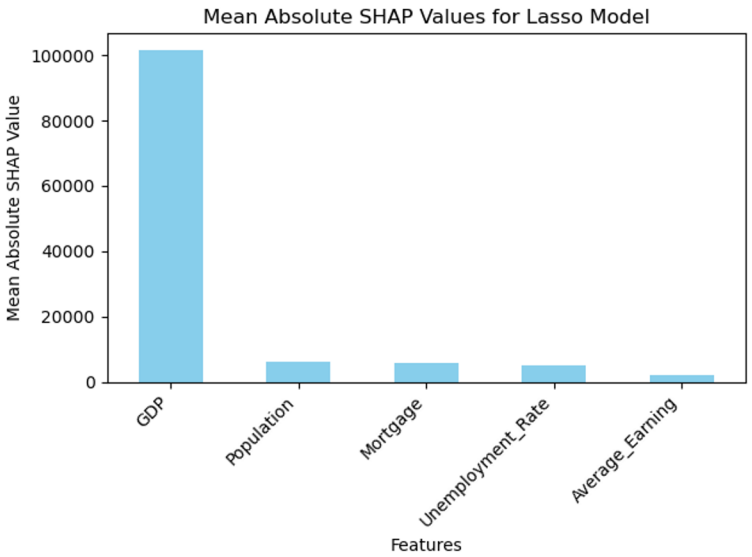

Fig.6. Shap_bar_plot_Lasso (Photo/Picture credit: Original)

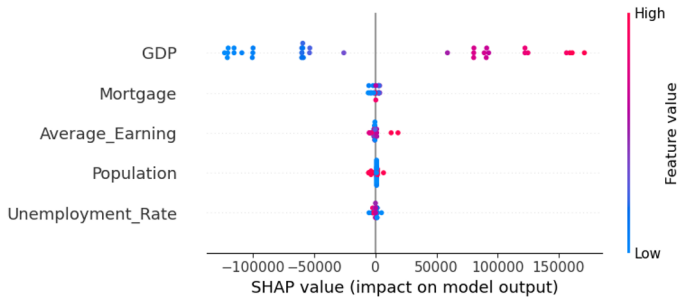

Fig.7. Shap_summary_plot_XGBoost (Photo/Picture credit: Original)

In the Lasso model, GDP emerged as the most significant feature, having the highest SHAP value. This underscores the pivotal role that GDP plays in influencing house prices within this model. In contrast, in the XGBoost model, GDP remained the most influential feature, followed by Mortgage and Average_Earning. This indicates that in a complex nonlinear framework, GDP is a key driver in house price prediction, while other variables like Mortgage and Average_Earning provide supplementary contributions. Features such as Population, Mortgage, and Unemployment_Rate had less impact in the Lasso model. However, XGBoost effectively captured more intricate nonlinear interactions, with GDP still being the primary feature, complemented by Mortgage and Average_Earning.

This study demonstrated that both the Lasso and XGBoost models performed well in predicting house prices. However, the models exhibited some discrepancies when forecasting high-priced properties, mainly due to the smaller sample size of such properties, leading to suboptimal fitting. Future research could incorporate additional variables, such as the internal amenities of houses and the socioeconomic conditions of the area, to enhance prediction accuracy. Furthermore, integrating interpretive tools like LIME could offer a clearer explanation of the model's decision-making process, thereby improving the model's interpretability and practical application [10].

4 Conclusion

This paper sought to enhance the accuracy of housing price prediction by comparing and analyzing the performance of the Lasso regression model and the XGBoost model. It provides both theoretical insights and empirical evidence for model selection in practical applications. The research utilized Bayesian optimization techniques to tune the hyperparameters of both models, in conjunction with data preprocessing and feature engineering, to optimize the overall predictive performance. Throughout the study, the Lasso model capitalized on its ability to perform feature selection, effectively reducing model complexity and mitigating the influence of irrelevant features. In contrast, the XGBoost model demonstrated its superior ability to model nonlinear relationships, particularly in scenarios involving complex feature interactions and high-value property data.

Analysis of key performance metrics such as MSE, R^2, and cross-validation error reveals that the optimized XGBoost model outperforms the Lasso model in terms of both accuracy and robustness. The XGBoost model is more effective when handling high-dimensional and nonlinear data. Furthermore, the data visualization analysis highlights the significant impact of key variables, such as GDP and average income, on housing prices. These findings provide valuable scientific insights for price prediction in the real estate market and establish a theoretical foundation for model selection and hyperparameter optimization.

In future research, the construction of a more comprehensive housing price prediction model could benefit from incorporating multidimensional factors such as regional economic

indicators and social infrastructure. Moreover, exploring the applicability of deep learning models in this domain could offer solutions to more complex prediction scenarios. This study provides a meaningful reference for real estate price prediction, with significant theoretical and practical implications for risk management, policy development, and investment decision-making within the real estate market.

References

1. A. A. Pilehvar, A. Ghasemi, Advanced modeling of housing locations in the city of Tehran using machine learning and data mining techniques, Humanities Soc. Sci. Commun. **11**, 804 (2024)
2. C. Monti, M. Pangallo, G. De Francisci Morales, F. Bonchi, On learning agent-based models from data, Sci. Rep. **13**, 9268 (2023)
3. Q. Truong, M. Nguyen, H. Dang, B. Mei, Housing price prediction via improved machine learning techniques, Procedia Comput. Sci. **174**, 433-442 (2020)
4. N. Vineeth, M. Ayyappa, B. Bharathi, House price prediction using machine learning algorithms, in I. Zelinka (Ed.), 2018 International Conference on Intelligent Systems, Springer, 425-433 (2018)
5. X. Xu, Y. Zhang, Residential housing price index forecasting via neural networks, Neural Comput. Appl. **34**, 14763–14776 (2022)
6. N. H. Zulkifley, S. Abdul Rahman, N. H. Ubaidullah, House price prediction using a machine learning model: A survey of the literature, I.J. Modern Educ. Comput. Sci. **6**, 46-54 (2020)
7. A. P. Singh, K. Rastogi, S. Rajpoot, House price prediction using machine learning, 2021 3rd International Conference on Advances in Computing, Communication Control and Networking, IEEE, 203-210 (2021)
8. R.-T. Mora-Garcia, M.-F. Cespedes-Lopez, V. R. Perez-Sanchez, Housing price prediction using machine learning algorithms in COVID-19 times, Land **11**, 2100 (2022)
9. W. K. O. Ho, B.-S. Tang, S. W. Wong, Predicting property prices with machine learning algorithms, J. Prop. Res. **38**(1), 48-70 (2020)
10. G. N. Satish, C. V. Raghavendran, M. D. Sugnana Rao, C. Srinivasulu, House price prediction using machine learning, Int. J. Innov. Technol. Explor. Eng. **8**(9), 717–721 (2019)