A Robust Ensemble-Based Framework for House Price Estimation: Integrating XG-Boost with SHAP and Web Deployment

Ezil Sam Leni. A[1*], *Revathi.* T[1], and *Sridhar* Devarajan[1]

[1]Alliance School of Advanced Computing, Alliance University, Bengaluru, India

Abstract. Accurate house price prediction is crucial for stakeholders in the real estate industry, including buyers, sellers, agents, and investors. Traditional valuation methods often rely on manual appraisal or basic regression techniques, which may lack scalability and the ability to capture complex relationships among features. With the rise of machine learning, especially ensemble models like XGBoost, there is a growing opportunity to improve prediction accuracy and robustness. This study presents a comprehensive framework for optimal house price prediction using XG-Boost, applied to the King County housing dataset comprising over 21,000 records. The suggested objective is supported by carefully conducted data preprocessing, which will imply outliers' elimination, feature engineering, and normalization. Other important derived features that were included were the age of the house and the renovation status of the house as well as the price per square foot. The XG-Boost regression model was trained and validated, and the R-squared ratio was found to be 0.87, indicating decent model performance. The model was also deployed as an interactive web application with Streamlit to allow users to store property details and get real-time predictions of the properties. Also, Shapley Additive Explanations (SHAP) values were applied to interpret the model output in a way that enhances explanations and builds trust among users. The system is not only accurate, but also keen on accountability, and usability. The proposed research will fill this missing gap and provide an actionable machine learning solution, deployable and explainable to real estate businesses to take advantage of the predictive analytics revolution. The results demonstrate the effectiveness of ensemble learning in conjunction with visualization and interpretability tools in the development of robust decision-support systems in the real-world housing markets.

1 Introduction

Most accurate house price prediction has remained an important aspect in the real estate business, which has touched prices of the involved stakeholders, including home buyers, home sellers, investors, financial institutions, and policymakers. In the past, real estate valuation of residential properties was done manually using some experience and knowledge, as well as a comparison of the market properties. These techniques, however, tend to be subjective, waste time, and scale poorly. As machine learning (ML) [1] technologies develop and the amount of data lies at our disposal, today, data-driven models provide potent substitutes to price estimation.

Real estate data, especially data in urban areas, is highly heterogeneous with many characteristics, including structural, such as bedrooms, bathrooms, and square footage; temporal, like year built and renovations; and location, including zipcode and proximity to amenities.

The interaction between these variables is normally not linear and many a time, tends to be complex [2].

Over the past few years, the Random Forests, Gradient Boosting Machines, and Support Vector Regressors machine learning models have been used to capture these patterns. Extreme Gradient Boosting (XGBoost) has become one of the better algorithms because of its regularization capabilities, scalability, and the fact that it can essentially handle missing data and multicollinearity reasonably well [1], [11]. The task implemented in the presented paper suggests the optimal model of the house price prediction based on the XG-Boost regressor trained on the dataset of King County, Washington, in the amount of more than 21,000 property records. The data provides rich material on properties describing their intricate features and their prices that have sold, making it possible to learn very well about complex pricing dynamics. The model is also integrated into an interactive web application via Streamlit, which makes it easy to perform and deploy. The latest research indicates the resilience of XGBoost in structured data issues. Studies reported by [2] showed that the prediction of XGBoost was better than Ridge Regression and even Decision Trees over the Root Mean Squared Error (RMSE) in a sample of urban housing data. On the same note, Troung indicated that in metropolitan Chinese cities, XG-Boost outperformed traditional models by dealing with heterogeneous house characteristics as well as non-linear interactions in

[*] Corresponding author: lenisatish@gmail.com

houses. Furthermore, Ho [3] also highlighted the decipherable feature and inter-market operability of XGBoost on geographically heterogeneous data.

The main objectives of this study are:

- To implement a predictive regression model of house prices with XG-Boost, taking advantage of the real-life dataset containing several features.
- To clean, impute, remove outliers and encode features to preprocess data successfully.
- To analyze the performance of the model with the help of the statistical measures like R^2, MAE, and RMSE.
- To turn the trained model into an operational web application on Streamlit to use in real-time.
- To give information on contributions of features with the help of the SHAP[8]-based method of interpretability.

Through achieving these goals, this study addresses the need for technical machine learning methods to have full-scale, realistic applications in the real estate realm.

2 Literature Survey

Machine learning has developed a useful property in recent years, as far as real estate analysis is concerned, specifically, the house price prediction (HPP). Sharma et al. [1] developed an XG-Boost-based prediction model specific to urban markets, considering engineered features such as near to public transport and neighborhood grouping. In their work, XG-Boost performed better than Ridge Regression and Decision Tree Regressors with better RMSE and R2 scores. Equally, Soltani [4] fed the XG-Boost algorithm with data on Chinese metropolitan regions and achieved improved results relative to linear regression and SVR. The researchers found that XG-Boost represents non-linear effects and local price variance better.

Adetunji [13] went further to show the efficiency of XG-Boost in a large-scale data set to include more than 20 features of a property; it was observed that XG-Boost was far more efficient and accurate than the traditional models. In this model, a sophisticated preprocessing is used that included label encoding and outlier detection capacity, which made his predictions more stable across each housing type. Li H. [5] conducted another study to compare XG-Boost and Decision Trees on national real estate listings. The results confirmed that XG-Boost has greater predictive power and is less sensitive to noisy data. Chen et al. [6] conducted a comparative analysis between XG-Boost and LightGBM, concluding that although LightGBM trains faster, XG-Boost yields higher accuracy due to its effective handling of missing data and regularization techniques. They highlighted the importance of tuning hyperparameters and integrating spatial and economic variables to maximize predictive performance. Fang [7] emphasized the use of SHAP for model interpretability, which revealed that features such as lot size and distance to the city center were among the top contributors to pricing. This underscores the growing emphasis not only on model accuracy but also on interpretability, an essential factor for end-user trust.

From a deployment perspective, Kamal et al. [12] introduced a full-stack ML pipeline combining XG-Boost with Streamlit to create an interactive HPP app. Their model allowed real-time predictions with user-friendly input controls and displayed model results via visual dashboards. Their cross-national study—spanning Indian and U.S. housing datasets—confirmed the generalizability of XG-Boost across diverse geographies.S. Im et al., [9] have developed an ensemble model for predicting the land prices with grip map data and SHAP. D. B. Acharya al., [10] demonstrated a framework for fairness and transparency in LightGBM and XGBoost models that can be applied to loan approval and house price prediction datasets.Current studies have also drawn attention to the necessity of presenting RMSE along with R^2 for interpreters of predictive models. For instance, Eze et al. [15] contrasted multiple regression with ensemble on a Boston Housing dataset and referred for comparison to RMSE as a performance index.

In summary, the reviewed literature confirms that XG-Boost consistently achieves high performance in housing price prediction tasks across regions and feature types. Combining XG-Boost with robust preprocessing, hyperparameter tuning, and deployment tools like Streamlit forms a reliable and scalable pipeline for real-world use cases. The present work draws upon these insights to construct an explainable, accurate, and deployable HPP model using King County data.

3 Research Methodology

This study follows a structured approach to develop a high-performing, interpretable, and deployable machine learning model for house price prediction using the XG-Boost regression algorithm. The overall methodology, shown in Figure 1, comprises four stages: data preprocessing, model training, web deployment, and model interpretability.

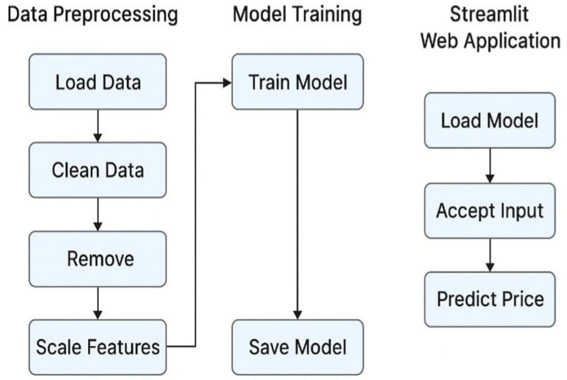

Fig. 1. System architecture of house price prediction.

3.1 Data Preprocessing

The dataset used for this study consists of 21,613 housing records from the King County, Washington Dataset [14]. Key features include ‘id’, ‘date’, ‘zipcode’, ‘lat’, ‘long’, ‘bedrooms’, ‘bathrooms’, ‘sqft_living’, ‘floors’, ‘condition,’ ‘grade,’ ‘waterfront,’ ‘view,’ ‘yr_built,’ and ‘yr_renovated’. In the data preparation process, at first, missing values were handled using median imputation. The price distribution reveals right-skewness, typical in real estate data. Outliers are removed using the interquartile range (IQR) method. Numerical features are normalized using StandardScaler to improve gradient-based model performance. The hyperparameters used in this research are summarized in Table 1.

Table 1. Abstract of hyperparameters in XGBoost.

Feature	Range
n_estimators	[100, 200]
learning_rate	[0.01, 0.1]
max_depth	[3, 5, 10]
subsample	[0.8, 1.0]
gamma	[0, 0.001]

3.2 Model Training Using XGBoost

The cleaned dataset is split into an 80:20 training and testing set. The XG-Boost Regressor, an ensemble learning model known for its robustness and accuracy with tabular data, is trained on the refined dataset. Hyperparameters was tuned to optimize model performance, and the resulting ensemble structure. The performance of the model is evaluated using the following metrics: coefficient of determination (R2) and mean squared error (MSE). The model training used in this research is as given below:

3.3 Web Deployment Using Streamlit

For accessibility, the model is integrated into a user-friendly interface built with Streamlit. Users can enter house attributes via sliders and inputs to receive real-time price predictions. The application internally preprocesses user input and forwards it to the trained XG-Boost model.

3.4 Model Interpretability and Visualization

To enhance transparency, the SHAP framework was employed to interpret predictions, displaying feature contributions. Feature importance scores were visualized to highlight influential variables. Additional visual aids such as heatmaps and scatter plots further supported user understanding of the model behavior.

4 Results and Discussion

In this research study, experimentation was performed on a standard computing environment with an Intel Core i7 processor, 16 GB RAM using Python 3.9 with libraries of XGBoost SHAP scikit-learn (v1.2.2), and Streamlit for deployment.

4.1 Feature Analysis

To provide interpretable insights into the proposed ensemble-based framework for house price estimation using the King County House Sales Dataset, SHAP is employed to analyse the feature contributions and understand the model's behaviour. SHAP assigns each feature an importance value for individual predictions, enabling both global and local interpretability. The top features influencing house prices in King County are summarized in Table 2.

Table 2. Feature analysis.

Rank	Feature	Importance Score
1	zipcode	571.0
2	lat	520.0
3	long	462.0
4	view	453.0
5	sqft_above	441.0
6	bathrooms	414.0

Table 3. Comparative analysis.

Model	MSE	RMSE	R² Score
XGBoost	1.933895×10^{10}	138,963	**0.8722**
Random Forest	2.217030×10^{10}	148,856	0.8535
Linear Regression	4.522487×10^{10}	212,719	0.7012
Decision Tree	4.487145×10^{10}	211,954	0.6771

4.2 Comparative Analysis

In regression analysis, let $y_i$denote the actual output and $\breve{y}_\imath$be the predicted output. Mean Squared Error (MSE) is a commonly employed metric to assess the average squared difference between actual and predicted values, as given in Equation (1). It evaluates model accuracy by computing the mean of squared residuals, with the squaring operation amplifying larger errors. Root Mean Square Error a standard metric for measuring regression models. It is the average of the absolute value of the errors but it doesn't provide the kind of relative decision making that MSE does, as it does not square the differences as expressed in Equation (2). The RMSE was calculated to understand the prediction errors in the dollar amount. XGBoost model with the lowest RMSE $138,963, outperforms Random Forest ($148,856), Linear Regression ($212,719) and Decision Tree (211,954). This further validates the better generality and practical feasibility of XGBoost for predicting real world housing prices. R-squared, or the coefficient of determination, is a statistical measure that quantifies how well a regression model captures the variability of the target variable. It indicates the proportion of variance in the dependent variable explained by the independent variables, as expressed in Equation (3).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \breve{y}_\imath)^2 \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \breve{y}_\imath)^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \breve{y}_\imath)^2}{\sum_{i=1}^{n}(y_i - \overline{y_\imath})^2} \tag{3}$$

The performance of the implemented XG-Boost model is compared with that of the linear regression, decision tree, and random forest as depicted in Figure 2. Table 3. gives the obtained values during experimentation. On evaluation, the XG-Boost model had an R² score of 0.872 and an MSE of 1.93 × 10^{10}, which are better than all the baseline models. In other words, the model explained 87.2 percent of the variance in house prices, indicating a fairly accurate model for baseline predictions. R² slightly outperformed Random Forest with 0.853, but neither Linear Regression (0.701) nor Decision Tree (677) showed as strong a performance as the others. This demonstrates the power of gradient boosting with regularization in capturing nonlinear feature interactions and reducing overfitting. The lower MSE of XG-Boost compared to all other models further confirms that XG-Boost has the best generalization for unseen test data.

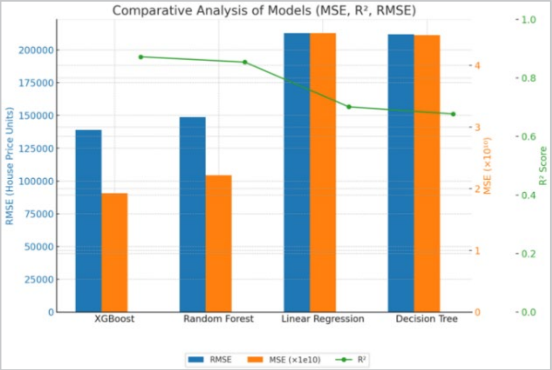

Fig. 2. Comparative analysis of MSE, R², RMSE.

4.3 Ablation Study

An ablation study involves systematically removing or modifying specific features from a predictive model to assess the importance of each feature in decision-making or forecasting. The result of the ablation study of this system is tabulated in Table 4 and captured in Figure 3. The removal of ‘zipcode’ resulted in the greatest loss of performance (R² from 0.872 to 0.858), It was tied with ‘sqft_above’ but in second place, which reminds us that location features are among the most important ones for house price prediction, and then lat.

Table 4. Performance- an ablation study.

Configuration	R²	MSE
Full model	0.872	1.93×10^{10}
zipcode	0.858	2.04×10^{10}
lat	0.861	2.01×10^{10}
long	0.864	1.99×10^{10}
view	0.866	1.97×10^{10}
sqft_above	0.861	2.02×10^{10}
bathrooms	0.867	1.96×10^{10}

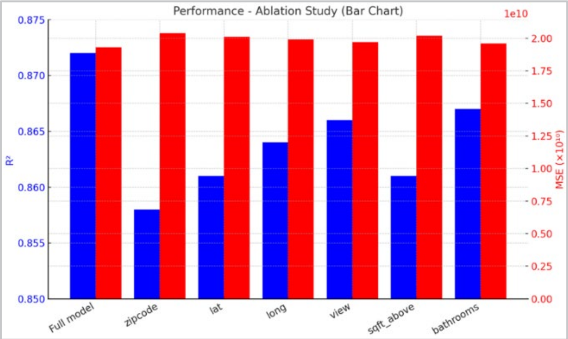

Fig. 3. Importance of features-an ablation study.

4.4 SHAP-Based Model Interpretability

SHAP was used to interpret the XG-Boost model predictions to maintain transparency. It has been demonstrated that SHAP values respect local accuracy and consistency since they are based on cooperative game theory and explain the feature impact to the difference between an average model prediction and a particular example output. Both a global and a local explanation were obtained using the TreeSHAP technique. It was not surprising that the most significant elements in the global SHAP study had to do with location (zipcode, lat, long), as this is one of the primary drivers of real estate values. Several other prominent features, like view, square footage above, and bathrooms, also had a significant impact. Local SHAP explanations for specific properties also shown how expected values were influenced by both positive (such delicious views and grade premium) and negative (like age of home creep) factors. In addition to providing intuitive information for the stakeholders, this duality of interpretability reinforced our earlier gain-based feature importance results. This enhanced confidence in the model's output and aligned with explainable AI concepts in a high-stakes industry like real estate.

4.5 Prediction Analysis

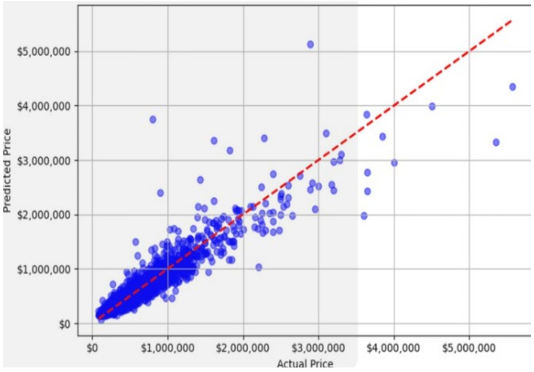

Fig. 4. Actual vs predicted house price.

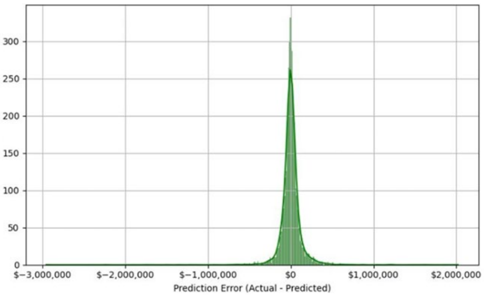

Fig. 5. Distribution of prediction errors.

Figure 4 shows a scatter plot comparing actual vs. predicted values. The close clustering of points along the diagonal confirms that the model closely matches ground truth prices, except for a few high-end outliers. The distribution of prediction error is shown in Figure 5, and the distribution is centered near zero and follows a bell-shaped curve, suggesting the model is unbiased. Most predictions fall within a ±$50, 000 error margin, which is acceptable for real estate applications with wide price ranges.

5 Conclusion

This study proposes a complete, practical, and interpretable solution for house price prediction using XGBoost, emphasizing both technical performance and user accessibility. By leveraging the King County housing dataset and applying robust preprocessing strategies, the model effectively captures key factors influencing property prices, including structural attributes, and living area. XG-Boost, known for its gradient boosting capabilities and regularization strengths, delivered superior accuracy compared to conventional approaches. As it comes to the test dataset, the archived R^2 value of 0.87 and low MSE, values validate the effectiveness of the model to learn complex connections between input feature attributes. SHAP analysis introduces an essential aspect of interpretability that can inform how each feature impacts the prediction and thus creates value by providing an accessible form of trustworthy and usable insights to non-technical stakeholders. In addition, the trained model deployment to a Streamlit-based web application proves the viability of its application in practice. End-users can use this interface to carry out interactions with the model easily and, therefore, estimate prices of houses in a reliable and simplified manner by individuals who are not particularly savvy in economic issues, specialists in the field of real estate and policy experts. To sum up, this work demonstrates that integrating ensemble of learning algorithms with explainable AI approaches and user-focused servicing platforms.

References

1. H. Sharma, H. Harsora, B. Ogunleye, An optimal house price prediction algorithm: Xgboost. Analytics **3**, 30-45 (2024). https://doi.org/10.3390/analytics3010003
2. 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). https://doi.org/10.3390/land11112100
3. W.K. Ho, B.-S. Tang, S.W. Wong, Predicting property prices with machine learning algorithms. J. Prop. Res. **38**, 48-70 (2021). https://doi.org/10.1080/09599916.2020.1832558
4. A. Soltani, M. Heydari, F. Aghaei, C.J. Pettit, Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. Cities **131**, 103941 (2022). https://doi.org/10.1016/j.cities.2022.103941
5. H. Li, House price prediction based on machine learning. Appl. Comput. Eng. **4**, 623-628 (2023). https://doi.org/10.54254/2755-2721/4/2023362
6. H. Chen, H. Jin, L. Li, Analysis and comparison of house price prediction based on XGBoost and LightGBM. Adv. Econ. Manag. Polit. Sci. **46**, 55-61 (2023). https://doi.org/10.54254/2754-1169/46/20230317Y
7. J. Fang, Forecast of foreclosure property market trends during the epidemic based on GA-BP neural network. Sci. Program. **2022**, 3220986 (2022). https://doi.org/10.1155/2022/3220986
8. S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. **30**, (2017). https://doi.org/10.48550/arXiv.1705.07874
9. S. Im, K. Kim, G. Lee, H.-J. Lim, Development of a weighted average ensemble model for predicting officially assessed land prices using grid map data and SHAP. IEEE Access **13**, 96251-96260 (2025). https://doi.org/10.1109/ACCESS.2025.3574698
10. D.B. Acharya, B. Divya, K. Kuppan, Explainable and fair AI: Balancing performance in financial and real estate machine learning models. IEEE Access **12**, 154022-154034 (2024). https://doi.org/10.1109/ACCESS.2024.3484409
11. M. Sharma, D. Sharma, R. Burle, P. Patil, I. Joge, C. Puri, Predicting house price model: A comprehensive analysis with random forest and decision tree method, in Proceedings of the 3rd International Conference for Innovation in Technology (INOCON), IEEE, Bangalore, India, May 06 (2024), 1–6
12. N. Kamal, E. Chaturvedi, S. Gautam, S. Bhalla, House price prediction using machine learning, in Proceedings of Emerging Technologies in Data Mining and Information Security (IEMIS 2020), Vol. 3, Springer, Kolkata, India, May 5 (2021), 799-811.
13. A.B. Adetunji, O.N. Akande, F.A. Ajala, O. Oyewo, Y.F. Akande, G. Oluwadara, House price prediction using random forest machine learning technique, Procedia Comput. Sci. **199**, 806-813. https://doi.org/10.1016/j.procs.2022.01.100
14. House Sales in King County, USA, Public Dataset, Kaggle (2016).
15. E. Eze, S. Sujith, M.S. Sharif, W. Elmedany, A comparative study for predicting house price based on machine learning, in Proceedings of the 4th International Conference on Data Analytics for Business and Industry (ICDABI), IEEE, Sakhir, Bahrain, August 15 (2024), 75-81