House Price Prediction Using Regression in Machine Learning

G. Veena Madhuri, A. Ajay Kumar, Md. Asrar Hussain, NSSS Girish Kumar, M.A.Jabbar

Department of CSE (AI&ML), Vardhaman College of Engineering, Hyderabad, Telangana

E-mail : veenamadhuri2516@gmail.com, ajayananthula0311@gmail.com, asrarhussain012@gmail.com, girishkumar.nistala@gmail.com, jabbar.meerja@gmail.com

2025 International Conference on Frontier Technologies and Solutions (ICFTS) | 979-8-3315-0276-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICFTS62006.2025.11031771

Abstract- The house pricing projection is crucial to the financial services industry and real estate sector. People are researching market techniques and attempting to purchase a new property within their means. However, the primary drawback of the current method is that it does not determine a home's price based on the essential forecast of future market trends, which results in an increase in price. Therefore, our paper's goal is to accurately forecast a home's price with minimal loss.For predicting house prices several factors need to be considered including location, size, amenities and economic indicators. There are very great techniques that machine learning offers for modelling and house price prediction. Regression algorithms are really valuable for predicting continuous values, like house prices by computing the association between data features and target values. The base tool in this direction is simply a basic linear approach, which fits best when discovering simple, linear trends. Decision trees and random forests add flexibility for modelling much more complicated and non-linear relationships. More advanced techniques like gradient boosting and XGBoost improve the accuracy in computing multiple models together to correct each other's errors iteratively. All these methods together provide a set of approaches for solving tasks in the price prediction problem under various accuracy and complexity requirements.

Keywords—Regression, Machine learning, Gradient Boosting, real estate, data preprocessing, feature selection, Mean Squared Error (MSE), R² score, hyperparameter tuning, feature importance, XGBoost.

I. INTRODUCTION

The real estate market is always complicated, comprising many factors such as the condition of the economy, location specific trends, and other variables that determine property characteristics. Traditional methods rely entirely on historical data and human judgment, often falling short when dealing with the complexities of real-world markets. In contrast, machine learning can process vast amounts of data to find patterns and trends that traditional analysis might miss. [1] [2] [3] [8]. The capacity of machine learning to offer more precise, data-driven insights into house price estimates is what inspired this work. Machine learning models can streamline real estate decision-making by utilizing sophisticated algorithms [4][5]. The study offers an accurate and effective housing price forecast model by utilizing cutting-edge machine learning techniques. This paper's primary contributions are as follows:

Model Development:
This paper introduces a strong learning model using XGBoost and places specific focus on real estate data. By comparing multiple regression models, including Gradient Boosting, the paper identifies the best-performing approach for house price prediction [1][4][7].

Data Preprocessing:
The model ensures proper handling of missing values, encoding of categorical variables, and normalization of features, which are critical steps to improve data quality and model reliability [8][9][11].

Streamlit-Based Application:
A user-friendly Streamlit application is developed, enabling users to input property features and obtain real-time house price predictions [11].

Model Evaluation and Deployment:
Detailed performance analysis using metrics such as Mean
Squared Error (MSE) and R² Score evaluates the reliability of the model. Additionally, the paper explores scalable deployment platforms for practical use in the real estate market [1][10].

Algorithm Comparison:
The paper examines the strengths and weaknesses of different algorithms, particularly XGBoost and Gradient Boosting, offering valuable insights into their applicability for real estate price forecasting [3][5][6]. This work represents a significant advancement in applying machine learning to predict realworld house prices, providing stakeholders in the real estate sector with a practical tool to make data-informed decisions [3][7][10].

II. Theoretical Background

This chapter compares the most important papers that have explored various methodologies for house price prediction. The most elementary method commonly used in house price prediction is linear regression, which serves as a basic yet user friendly approach. Linear regression models the relationship between house prices and their features, such as location, house size, and market trends. While it is easy to implement and interpret, research has shown that linear regression struggles to capture the non-linear relationships that often exist between these features and house prices [1][2][3]. Table 1 summarizes key studies in this area, highlighting the techniques used, achieved accuracy, and the main methodology of each work.

Table 1 Different Ml/Dl Techniques Used For House Price Prediction

Ref No	Techniques used	Methodology	Accuracy
[1]	Deep Learning, Self-Attention	Deep learning model with heterogeneous data analysis and joint self-attention mechanism	90%
[2]	Neural Networks, Machine Learning	Combines neural networkswith traditional ML approaches to improve predictive accuracy	92%
[3]	Support Vector Machine	Applies various machine learning algorithms based on local property data	88%
[4]	CNN, Adaboost	UtilizesML techniques to determine housing price trends	94%
[5]	Regression Techniques	Employs feature engineering and regression analysis to build a predictive pricing model	89%
[6]	Comparative Regression Analysis	Evaluates and compares different regression approaches for optimal price prediction	91%
[7]	Linear regression	Reviews and assesses the applicability of machine learning models in real estate	85%
[8]	Different ML and DL algorithms	ML and DL methods for house price prediction	90%
[9]	Feature selection, ML	Significant features analysis for house price prediction using machine learning	93%

In summary, the majority of recent papers confirm that XGBoost offers the best balance between accuracy and computational efficiency, especially when applied to large and complex real estate datasets. This paper builds on these insights by implementing XGBoost alongside Gradient Boosting for house price prediction, offering a detailed comparison and real-world application through a Stream lit interface.

III. Machine Learning Techniques

Machine Learning (ML) algorithms analyse data to find trends, predict the price of the house, and help cost minimization [1][2][3][11]. This contributes significantly to the success of real estate market. Some popular machine learning methods for prediction are listed below:

A. Linear Regression
Linear regression is a statistical method that models the relationship between a dependent variable (house price) and one or more independent variables (features like location, size, and number of rooms). It assumes a linear relationship between the input features and the output [3][4][5][6].

- Usage in House Price Prediction:Useful for simple cases where the relationship between variables is approximately linear, such as predicting house price based on square footage.
- Pros: Easy to implement and interpret, especially for smaller datasets.

- Cons: It struggles with non-linear relationships and can be sensitive to outliers.

B. Decision Tree

Decision trees create a corresponding tree structure while dividing the data into smaller sections. Features are represented by the decision nodes, and the expected output house price is represented by the leaves. The feature that yields the highest information gain or the greatest variance reduction is the basis for each split [3][4][8].

- Usage in House Price Prediction: Can model complex relationships and capture interactions between features like location, age of the house, and number of rooms.
- Pros: Easy to interpret, can handle both numerical and categorical features, and captures non-linear relationships.
- cons: Prone to overfitting if not properly pruned and may perform poorly on unseen data.

C. Bagging (Random Forest)

In order to increase accuracy and decrease overfitting, Random Forest is an ensemble technique that constructs several decision trees and aggregates their predictions. A random subset of the characteristics and data is used to train each tree.

- Usage in House Price Prediction: Highly effective for capturing non-linear relationships and complex feature interactions, such as amenities, proximity to schools, and public transport.
- Pros: Reduces overfitting, handles missing data well, and can rank the importance of features.
- Cons: Requires more computational resources and is less interpretable compared to single decision trees.

D. Gradient Boosting Machines (GBM)

Sequentially, GBM constructs a group of decision trees, each of which aims to fix the mistakes of the one before it.The predictions are improved gradually over iterations [4] [5[6].

- Usage in House Price Prediction: Excellent for handling complex relationships and large datasets where small improvements in accuracy are crucial, like predicting price variations in luxury properties.
- Pros: High accuracy, can handle both classification and regression tasks, and good at reducing bias.
- Cons: Computationally expensive, especially for large datasets, and prone to overfitting without careful tuning.

E. XGBoost (Extreme Gradient Boosting)

An enhanced variant of gradient boosting is called XGBoost, designed for performance and speed. It includes features like parallel processing, regularization to prevent overfitting, and advanced handling of missing data.

- Usage in House Price Prediction: Frequently used for real-world house price prediction problems due to its ability to handle large datasets and its high accuracy.
- Pros: Fast, accurate, handles missing data well, and includes built-in regularization to prevent overfitting.
- Cons: Requires careful hyperparameter tuning, and training can be computationally intensive.

F. Support Vector Machines (SVM)

SVM is helpful for both classification and regression tasks because it looks for the hyperplane in feature space that optimally divides the data points. SVM is utilized for regression (SVR) in the context of predicting home prices [3][4].

- Usage in House Price Prediction: Can be useful when there are non-linear relationships in the data, such as when house prices are affected by unique or rare features.
- Pros: Effective for high-dimensional datasets, can handle non-linear relationships using kernels.
- Cons: Computationally expensive, less interpretable, and sensitive to feature scaling.

G. K-Nearest Neighbours (KNN)

KNN is an easy, instance-based learning algorithm that makes predictions by averaging the prices of the k-nearest data points (houses) in the feature space [1][3].

- Usage in House Price Prediction: Works well when the price of a house is likely to be influenced by its proximity to similar houses in terms of features like location and size.
- Pros: Easy to understand and implement, no training time required.
- Cons: Costly to compute at prediction time, has trouble with massive data sets, and works in extremely dimensional areas.

IV. Performance Analysis

To assess the performance and effectiveness of the Gradient Boosting and XGBoost regression models in predicting house prices, a series of experiments were conducted. The results of these experiments provide valuable insights into the accuracy and reliability of each model. The key metrics used for evaluation include Mean Squared Error (MSE) and R^2 score. [2][5][7].

A. Gradient Boosting Regressor

The MSE value indicates the average squared difference between the actual house prices and the predicted prices. A lower MSE value signifies better model performance. The Gradient Boosting Regressor's MSE is relatively high, suggesting room for improvement in prediction accuracy [5][6].

The R^2 score represents the proportion of the variance in house prices that can be explained by the model. An R^2 score closer to 1 indicates a better fit. The Gradient Boosting Regressor achieved an R^2 score of approximately 0.61, indicating that it explains about 61% of the variance in house prices based on the input features.[11][12]

B. XGBoost Regressor

The XGBoost Regressor achieved a lower MSE compared to the Gradient Boosting Regressor. This lower MSE value demonstrates that the XGBoost model makes more accurate predictions, reducing the average squared difference between actual and predicted house prices [3] [4]. The R^2 score for the XGBoost Regressor is significantly higher at approximately 0.64. This indicates that the XGBoost model explains about 65% of the variance in house prices, showcasing its superior ability to capture the underlying patterns in the data [7].

C. Comparative Analysis

Exploratory Data Analysis (EDA) was conducted on the house price prediction dataset using Google Colab to understand the data distribution and identify key patterns. Following EDA, regression models, including Gradient Boosting Regressor and XGBoost Regressor, were trained and evaluated based on performance metrics such as Mean Squared Error (MSE) and R^2 score. The performance comparison was visualized through bar plots, illustrating the models' effectiveness. These visualizations were generated and extracted directly from the Google Colab environment. The comparative analysis of the two models highlights the superiority of the XGBoost Regressor in terms of both MSE and R^2 score. Specifically, the XGBoost Regressor's MSE is substantially lower than that of the Gradient Boosting Regressor, demonstrating its enhanced accuracy in price prediction. The higher R^2 score of the XGBoost Regressor indicates that it provides a better fit to the data, capturing more of the variance in house prices compared to the Gradient Boosting Regressor.

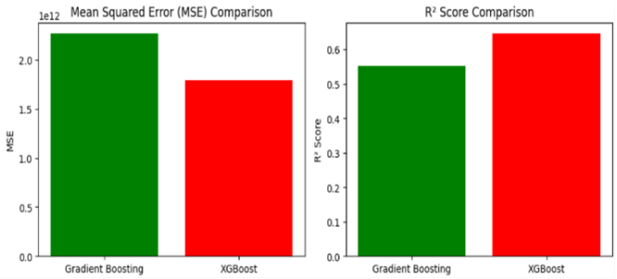

Fig. 1. Model Performance Comparison

For the Gradient Boosting Regressor, the scatter plot shows a wider spread around the ideal diagonal line (where actual prices equal predicted prices), indicating less accurate predictions.For the XGBoost Regressor, the scatter plot points are more tightly clustered around the ideal diagonal line, illustrating more accurateand reliable predictions.

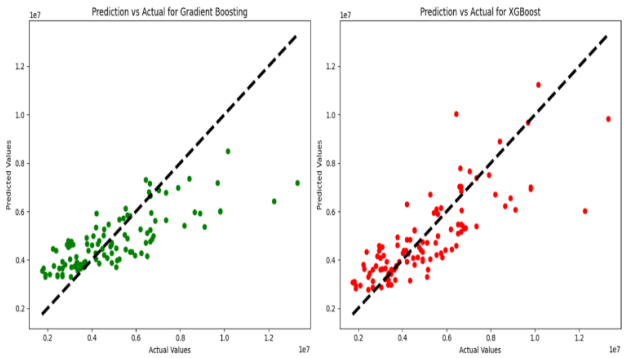

Fig. 2. Scatter plots of actual vs predicted prices

The residual plots (actual price - predicted price) for both models show the distribution of errors. The XGBoost Regressor's residual plot has smaller and more uniformly distributed residuals, indicating fewer and less significant prediction errors compared to the Gradient Boosting Regressor.

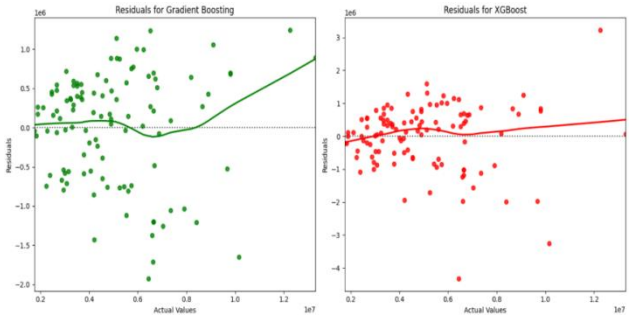

Fig. 3. Residual plots

These results clearly show that the XGBoost model outperforms the Gradient Boosting model, making it the preferred choice for house price prediction in this study.

V. XGBoost (Extreme Gradient Boosting)

XGBoost is an advanced machine learning algorithm specifically optimized for speed and performance. It's based on the Gradient Boosting framework, but with several improvements that make it particularly effective for predictive tasks like house price prediction. XGBoost builds an ensemble of decision trees sequentially, each correcting the errors of the previous one [4][11][12].

A. How XGBoost Works

- Gradient Boosting Foundation: XGBoost builds trees sequentially by minimizing the errors of previous trees. This iterative approach helps in learning complex patterns in the data, making it effective for tasks with a mix of numerical and categorical features, such as house price prediction [6].
- Optimized Regularization: XGBoost adds L1 (Lasso) and L2 (Ridge) regularization, which helps prevent overfitting by penalizing complex models. This is useful in-house price prediction, where the model could otherwise fit to noise in the training data [2][3].
- Handling Missing Values: XGBoost has a built-in capability to handle missing data, which is common in real estate data (e.g., missing values for amenities or property features). The algorithm assigns default directions in trees for missing values, allowing it to still make predictions without additional preprocessing [2][7][9].
- *Parallel Processing:* XGBoost's design enables parallel tree construction, making it faster than traditional boosting methods. This is particularly advantageous for large datasets like those used in real estate, which might contain numerous features and entries [1][2][9].

B. Advantages of XGBoost for House Price Prediction

XGBoost's advanced features make it particularly suited for the complex and high-dimensional nature of real estate data. Here's why it often outperforms other algorithms:

- High Predictive Accuracy: XGBoost is designed to reduce both bias and variance, resulting in high predictive accuracy. In house price prediction, this means the model can better handle the complexities of factors like location, size, amenities, and economic indicators [6][11].
- Handles Complex Data Structures: House price data often involves non-linear interactions between variables (e.g., proximity to amenities vs. price), which XGBoost's boosted trees are better suited to capture compared to linear models or simpler algorithms like KNN or Decision Trees [1][8].
- Efficient for Large Datasets: Real estate datasets can be large, with numerous features describing property characteristics, neighbourhood details, and market conditions. XGBoost's parallel processing and efficient memory usage make it faster and more scalable than Gradient Boosting or Random Forests [1][2][7].
- Regularization Reduces Overfitting: Regularization (L1 and L2) in XGBoost prevents the model from fitting noise, which is common in real estate data due to fluctuating prices and outliers. This helps maintain model generalizability on new data [1][12].
- *Automated Feature Handling:* With its ability to handle missing values and automatically encode categorical variables, XGBoost reduces the amount of manual preprocessing needed for real estate data, streamlining the model training pipeline [1][8][10].

C. Comparison with Other Algorithms

- Linear Regression: Although useful for simpler, linear relationships, linear regression struggles with the non-linear and interaction-heavy nature of house price data, making it less accurate than XGBoost [3][4][12].
- Random Forest: While Random Forest handles nonlinear data well, it can be slower and less memory-efficient than XGBoost. Additionally,

Random Forests lack the regularization of XGBoost, making them more prone to overfitting on complex data [11].

- *Gradient Boosting Machines (GBM):* Traditional GBM is similar to XGBoost but lacks the speed and optimizations like parallel processing and missing value handling, which are essential for large-scale real estate datasets [6].
- *Support Vector Machines (SVM):* SVMs can be effective but are computationally intensive on large datasets and less interpretable compared to XGBoost. They also require significant feature scaling and tuning [2]10].

VI. Conclusion

After conducting a comprehensive analysis using the Housing.csv dataset, comparison was conducted between the performance of Gradient Boosting and XGBoost Regression models for predicting house prices. Our findings indicate that XGBoost consistently outperformed Gradient Boosting in terms of predictive accuracy and computational efficiency. Specifically, XGBoost demonstrated lower Root Mean Squared Error (RMSE) and higher R-squared values, suggesting its superior ability to capture complex relationships within the data. This performance advantage can be attributed to XGBoost's optimized tree boosting algorithm, which effectively handles large datasets and minimizes overfitting. Furthermore, our feature importance analysis in XGBoost revealed critical insights into the factors influencing house prices. Features such as location, square footage, and number of bedrooms emerged as significant predictors, underscoring their impact on property valuations. By understanding these key drivers, stakeholders can make more informed decisions in real estate investment and market analysis. In conclusion, the XGBoost model not only excels in predictive accuracy but also offers robustness and scalability for real-world applications. Its ability to generalize well to unseen data positions it as a powerful tool for housing market predictions, setting a benchmark in machine learning regression tasks.

References

[1]. Wang, Pei-Ying, Chiao-Ting Chen, Jain-Wun Su, Ting-Yun Wang, and Szu-Hao Huang. "Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism." IEEE access 9 (2021): 55244-55259.

[2]. Varma, Ayush, Abhijit Sarma, Sagar Doshi, and Rohini Nair. "House price prediction using machine learning and neural networks." In 2018 second international conference on inventive communication and computational technologies (ICICCT), pp. 1936-1939. IEEE, 2018.

[3]. Phan, The Danh. "Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia." In 2018 International conference on machine learning and data engineering (iCMLDE), pp. 35-42. IEEE, 2018.

[4]. Banerjee, Debanjan, and Suchibrota Dutta. "Predicting the housing price direction using machine learning techniques." In 2017 IEEE international conference on power, control, signals and instrumentation engineering (ICPCSI), pp. 2998-3000. IEEE, 2017.

[5]. Manasa, J., Radha 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), pp. 624-630. IEEE, 2020.

[6]. Madhuri, CH Raga, G. Anuradha, and M. Vani Pujitha. "House price prediction using regression techniques: A comparative study." In 2019 International conference on smart structures and systems (ICSSS), pp. 1-5. IEEE, 2019.

[7]. Singh, A.P., Rastogi, K. and Rajpoot, S., 2021, December. House price prediction using machine learning. In 2021 3rd international conference on advances in computing, communication control and networking (ICAC3N) (pp. 203-206). IEEE.

[8]. Zhou, Jincheng, Tao Hai, Ezinne C. Maxwell-Chigozie, Afolake Adedayo, Ying Chen, Celestine Iwendi, and Zakaria Boulouard. "Effective House Price Prediction Using Machine Learning." In International Conference on Advances in Communication Technology and Computer Engineering, pp. 425-436. Cham: Springer Nature Switzerland, 2023.

[9]. Saefudin, M.R., Putri, M.R., Hadi, A., Wijayanto, H. and Irmawati, B., 2024, November. Significant Features for House Price Prediction Using Machine Learning. In 2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT) (pp. 659-664). IEEE.

[10]. Peng, Hao, Jianxin Li, Zheng Wang, Renyu Yang, Mingsheng Liu, Mingming Zhang, S. Yu Philip, and Lifang He. "Lifelong property price prediction: A case study for the toronto real estate market." IEEE Transactions on Knowledge and Data Engineering 35, no. 3 (2021): 2765-2780.

[11]. Verma, Abhay, Durgesh Singh, Netra Patil, Sagar G. Mohite, Sakshi Ranjan, and Mohit Raj. "Enhanced House Price Prediction Using Machine Learning Techniques." In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-6. IEEE, 2024.

[12]. Singh, Anannya, Ajay Vikram Singh, and Sonu Mittal. "Housing price prediction using multiple regression." In Computational Methods in Science and Technology, pp. 433-441. CRC Press, 2024.