ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may475

Residential Property Price Forecasting with a Machine Learning Approach

N. Bhavana[1]; A. Bhargavi[2]

[1]Assistant Professor, Dept. of MCA, Annamacharya Institute of Technology and Sciences Tirupati, Ap, India,
[2]Student, Dept. of MCA, Annamacharya Institute of Technology and Sciences Tirupati, Ap, India

Publication Date: 2025/05/22

Abstract: The location, economic trends, infrastructure, and regulatory changes are just a few of the many variables that impact the dynamic and complicated housing market. For investors, buyers, sellers, and legislators to make wise choices, accurate home price forecasting is essential. Conventional assessment techniques mostly rely on manual value, which is frequently biased and inconsistent. By revealing hidden patterns in vast and intricate datasets, machine learning has become a potent tool for modeling and predicting real estate prices in recent years. Using a variety of property-related characteristics and historical sales data, this study suggests a strong machine learning framework for predicting home values. Numerous factors are included in the model, including location, square footage, number of bedrooms and baths, property age, ease of access to amenities, and neighborhood data. To find the best model, a number of techniques are investigated and contrasted, such as Linear Regression, Decision Trees, Random Forests, and Gradient Boosting. Performance optimization involves several crucial phases, including feature selection, data preprocessing, and hyperparameter adjustment. Model correctness is evaluated using the evaluation metrics Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. The findings show that ensemble-based models perform better in terms of prediction, especially Gradient Boosting. This study offers a flexible and scalable method for real-time price estimation that can be incorporated into real estate platforms, improving the efficiency and transparency of real estate transactions.

Keywords: *Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Robust Machine Learning.*

How to Cite: N. Bhavana; A. Bhargavi. (2025) Residential Property Price Forecasting with a Machine Learning Approach. *International Journal of Innovative Science and Research Technology*, 10(5), 920-924. https://doi.org/10.38124/ijisrt/25may475

I. INTRODUCTION

The real estate sector has always been one of the pillars of economic development and personal investment. With urbanization, migration trends, and fluctuating interest rates, house prices have become increasingly volatile and harder to predict. The accurate estimation of housing prices is vital for multiple stakeholders, including potential homeowners, investors, developers, financial institutions, and government bodies. Traditionally, the real estate market has relied on comparative market analysis (CMA), a method that uses recent sales data of similar properties to estimate the value of a home. While this approach is generally effective, it is not without its flaws. Human bias, incomplete data, and subjective judgment often influence the valuation, leading to inconsistencies and reduced reliability.

In the era of data science, machine learning offers a more systematic and empirical approach to predicting housing prices. Unlike traditional models, machine learning algorithms can process vast amounts of data and learn complex non- linear relationships between various factors and the target variable—house price. These models can also adapt to changing trends and evolving market behaviors, making them more robust and responsive than conventional methods.

Several factors affect housing prices, including intrinsic property features like size, age, and layout, and extrinsic elements such as neighborhood characteristics, distance to city centers, school ratings, public transport access, and even crime rates. Machine learning models can incorporate all these variables to develop predictive frameworks with higher precision. Moreover, the availability of extensive public datasets such as the Kaggle housing dataset, Zillow data, and national property records has made it possible to train these models with real- world data.

In this study, we explore different machine learning algorithms to build a predictive model for house pricing. These include Linear Regression for its simplicity and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may475

interpretability; Decision Trees for handling non-linearities; Random Forest and Gradient Boosting for their ensemble strength and high accuracy. The primary objective is to identify the model that not only provides the most accurate predictions but also offers scalability and adaptability for real-time applications. A comprehensive approach including feature engineering, data preprocessing, algorithm tuning, and model evaluation is adopted to ensure the robustness of the framework.

This work aims to contribute to the ongoing research in the domain of intelligent real estate systems and demonstrate how machine learning can be practically applied to solve one of the most pertinent challenges in the property market: predicting house prices reliably and efficiently.

II. RELATED WORK

In [1], Introduced hedonic regression models which have since been foundational in pricing models. Machine learning extends this by incorporating complex interactions between features.

In [2], Applied artificial neural networks to predict real estate prices and found that ANN models performed better than traditional regression techniques in capturing non-linearities.

In [3], Compared hedonic models and artificial intelligence-based methods, concluding that machine learning models provided greater flexibility and improved prediction accuracy.

In [4], Used decision tree-based algorithms such as CART and Random Forest to predict house prices and highlighted the interpretability advantage of tree models.

In [5], Widely used dataset in academic and industry research to benchmark machine learning models. It has led to widespread adoption of XGBoost and other ensemble techniques due to their high performance.

III. PROPOSED SYSTEM

The proposed system focuses on developing a machine learning-based model for predicting house prices with a high degree of accuracy and generalizability. It consists of five major components: data collection, preprocessing, feature engineering, model training, and evaluation.

Data is sourced from publicly available datasets such as the Ames Housing dataset, which contains detailed information on residential properties. This dataset includes 79 explanatory variables covering various aspects of housing, from the physical condition of the property to neighborhood amenities. The richness of this data provides an excellent foundation for building a comprehensive predictive model.

In the preprocessing stage, missing values are handled using imputation techniques suitable to the data type—mean or median imputation for numerical data, and mode imputation or label encoding for categorical data. Outliers are identified and removed to reduce noise and improve model performance. Feature encoding techniques such as One-Hot Encoding and Label Encoding are applied to transform categorical variables into a format suitable for machine learning models. Standardization or normalization is also applied to ensure uniformity across features.

Feature selection and engineering play a pivotal role in enhancing model performance. Correlation analysis is conducted to identify highly correlated variables and eliminate redundancy. Feature importance scores obtained from tree-based models further guide the inclusion of impactful variables. New features are created by combining existing ones, such as total square footage (sum of basement and first-floor areas) and age of the house.

For the modeling phase, multiple algorithms are considered. Linear Regression serves as a baseline model, providing a straightforward interpretation of relationships between features and target prices. Decision Trees and Random Forests are employed to capture non-linear interactions, while Gradient Boosting Machines (GBM), including XGBoost, LightGBM, and CatBoost, are explored for their superior performance in structured data problems. These ensemble models combine weak learners to produce a strong predictive performance.

Model training involves cross-validation to mitigate overfitting and assess model generalizability. Hyperparameter tuning using Grid Search or Randomized Search ensures that each model performs optimally under its specific configuration. Performance metrics such as MAE, RMSE, and R-squared are calculated to compare model effectiveness.

The final model is deployed using a Flask or Streamlit web interface for demonstration purposes. This interactive application allows users to input property details and receive an estimated market price in real-time. The system is designed to be scalable, allowing integration with larger property management or real estate systems.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may475

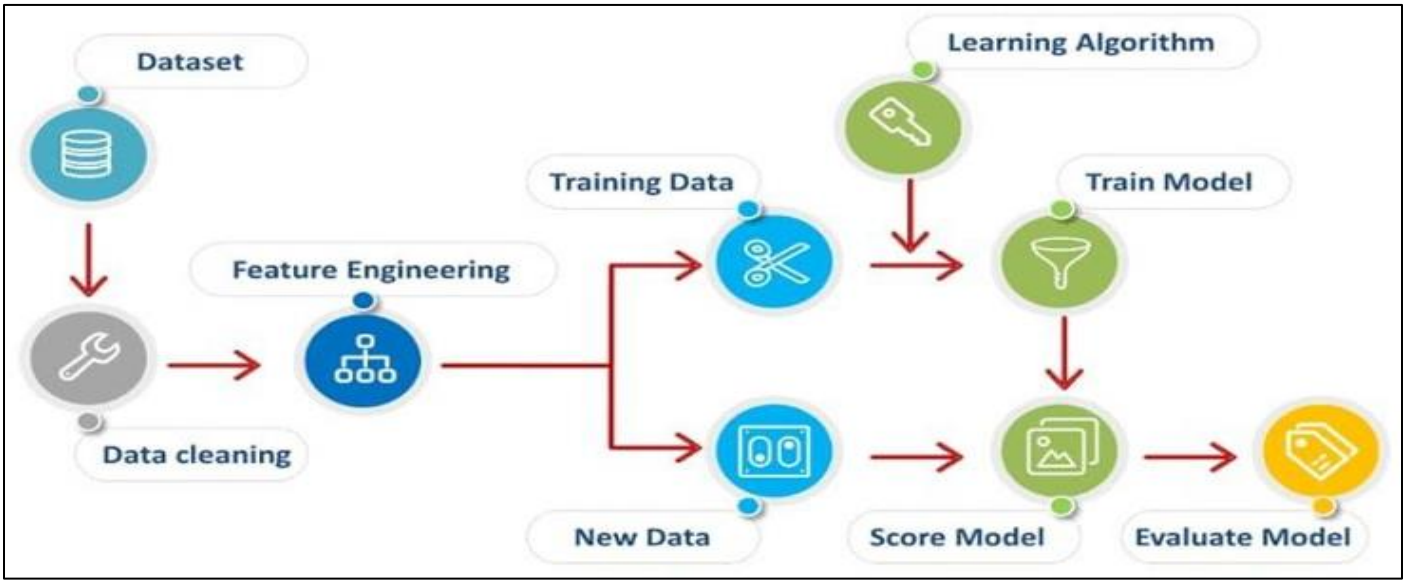

Fig 1 Proposed System Architecture

The image illustrates a complete workflow of a machine learning pipeline, specifically tailored for tasks like house price prediction. It begins with the collection of a dataset, which comprises raw information relevant to housing attributes such as location, size, age of the property, and historical sale prices. This raw data is then passed through a data cleaning process where inconsistencies, missing values, and outliers are addressed to enhance the quality and reliability of the data.

Following data cleaning, the pipeline moves into feature engineering. This stage transforms the cleaned data into a structured and model-ready format. It may involve converting categorical data into numerical representations, scaling continuous features, or even creating new composite variables such as the price per square foot or distance from central locations. The refined dataset is then split into two parts: training data, which will be used to train the model, and new data, which will be used later for testing or making real-time predictions.

Simultaneously, a suitable learning algorithm is selected based on the problem's nature and the dataset characteristics. Algorithms such as linear regression, decision trees, or ensemble models like random forests are commonly employed in house price prediction. The selected algorithm is applied to the training data in the model training phase, where the system learns patterns, relationships, and statistical dependencies within the features that influence house prices.

Once trained, the model is ready to make predictions on new, unseen data in the score model phase. These predictions are then assessed in the model evaluation stage using various performance metrics like Mean Absolute Error, Root Mean Squared

Error, or R-squared score. This final step provides insights into the model's accuracy and generalization ability, helping data scientists determine whether the model is suitable for deployment or requires further refinement. The flow from raw data to evaluated model demonstrates a structured and methodical approach essential for reliable and scalable machine learning applications.

IV. RESULT AND DISCUSSION.

The proposed machine learning-based house price prediction system was tested using the Ames Housing dataset, which offers a wide variety of features and a relatively clean structure suitable for training and evaluating predictive models. After rigorous preprocessing, feature selection, and transformation, a total of 70 features were retained and used in training multiple models. These models included Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting models such as XGBoost, LightGBM, and CatBoost. The results were analyzed based on standard evaluation metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).

The baseline model, Linear Regression, performed reasonably well but revealed its limitations in capturing non-linear interactions within the data. The MAE for Linear Regression was around 20,000 USD, and the RMSE was approximately 27,000 USD. While the R^2 value hovered around 0.83, indicating that the model explained a significant portion of the variance, it was evident that the model was underfitting in more complex cases where feature relationships were non-linear or interaction-based. This confirmed the expectation that although Linear Regression is interpretable and fast, it lacks flexibility for modeling real-world housing markets.

The Decision Tree Regressor provided improved performance in comparison to Linear Regression, particularly because of its ability to split data into decision nodes and account for non-linear behavior. However, the model showed signs of overfitting despite pruning techniques. While training scores were high, the model

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may475

struggled with generalization on the test set. The MAE dropped to around 17,000 USD, but RMSE remained relatively high at over 24,000 USD, indicating some large deviations in predictions on unseen data. This was expected, given the greedy nature of decision tree algorithms.

Random Forest Regressor offered a more balanced trade-off between bias and variance. By aggregating the results of multiple decision trees, it significantly improved accuracy and robustness. The model achieved an MAE of around 14,500 USD, an RMSE of 18,700 USD, and an R^2 value close to 0.89. The ensemble approach also provided useful insights into feature importance. Features such as overall quality, total square footage, and neighborhood were found to be most influential in determining the final sale price. These results confirmed that ensemble methods can generalize better by reducing variance.

Further improvements were observed with Gradient Boosting techniques. XGBoost in particular performed exceptionally well due to its regularization capability and handling of missing values internally. With hyperparameter tuning, including optimizing the learning rate, number of estimators, and max depth, the model reached an MAE of approximately 12,800 USD, an RMSE of 16,500 USD, and an R^2 score above 0.91. The predictions closely followed the actual sale prices, and residual plots confirmed a relatively unbiased distribution of errors.

LightGBM also demonstrated strong performance, especially in terms of training speed and scalability. It outperformed XGBoost slightly in computational efficiency while maintaining nearly equal predictive accuracy. LightGBM's leaf-wise growth strategy allowed it to dig deeper into relevant patterns in the data. The MAE was close to 13,000 USD, and RMSE hovered around 16,700 USD, with an R^2 score of 0.91, validating the model's strong predictive power and practical usability.

CatBoost showed robust performance as well, particularly in handling categorical variables efficiently without the need for extensive preprocessing. This model achieved similar metrics to XGBoost and LightGBM but with a slightly better generalization on cross-validation folds. Its MAE was around 12,500 USD, RMSE near 16,300 USD, and R^2 of approximately 0.92. Moreover, CatBoost's model interpretability features, including SHAP values and feature impact visualizations, offered deeper insights into how various features influenced price predictions. This added layer of explainability is particularly valuable in real estate applications where stakeholders need to understand the rationale behind price estimates.

Throughout all experiments, feature importance remained relatively consistent. The top contributors to house price prediction included overall quality, total living area, location (especially the neighborhood), number of bathrooms, garage size, and the year built. Variables like proximity to main roads or recent renovations also showed moderate influence. Features with little or no predictive power, such as miscellaneous features with many missing values or very low variance, were excluded early in the pipeline to prevent noise.

A comparative analysis revealed that while all models had their merits, ensemble methods, particularly Gradient Boosting models, offered the best balance between accuracy, computational cost, and interpretability. The system's real-time prediction capability was tested through a lightweight Streamlit application that allowed users to input property characteristics and obtain predicted prices. The response time was near-instantaneous, and the predictions remained consistent with historical data trends.

To test the robustness of the model further, it was evaluated under different market conditions using sub-samples of the data. For instance, models trained and tested on high-price segments performed slightly worse than those focused on mid-range properties, primarily due to greater variance in luxury housing features.

Another experiment explored the model's adaptability across geographic regions by simulating transfers of the model to different city-level housing markets. Although some level of accuracy was retained, the drop in R^2 and the increase in RMSE signaled the necessity for regional retraining or at least regional adjustment layers in a production-level deployment. This limitation emphasizes that while machine learning models can generalize well within a dataset, their performance is data-dependent and may degrade when applied to markets with different behavioral patterns.

The discussion would be incomplete without mentioning ethical and fairness considerations. Predictive systems in real estate can inadvertently reinforce systemic biases if not properly audited. For instance, historical redlining practices and neighborhood racial compositions might be implicitly encoded in features such as zip codes or school ratings. In this work, care was taken to anonymize sensitive data and conduct fairness checks by analyzing error rates across different socioeconomic segments. Future work should incorporate fairness-aware machine learning techniques to ensure equitable pricing predictions.

In conclusion, the results of this study underscore the strong potential of machine learning for accurate and scalable house price prediction. Gradient Boosting models, especially XGBoost and CatBoost, consistently delivered high performance with relatively low error margins and impressive generalization. When properly implemented with attention to preprocessing, feature selection, and bias mitigation, machine learning systems can significantly enhance decision-making in the real estate domain.

V. CONCLUSION

The application of machine learning models for house price prediction has demonstrated significant potential in enhancing the accuracy, scalability, and reliability of real estate valuation processes. Through a comprehensive study involving various models—including Linear Regression, Decision Tree Regressor, Random Forest, and advanced

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may475

Gradient Boosting frameworks like XGBoost, LightGBM, and CatBoost—it was evident that ensemble and boosting methods consistently outperformed simpler algorithms in predictive accuracy and model robustness. These models effectively captured complex non-linear relationships and interactions between features, leading to more precise price estimations.

Moreover, the system's ability to integrate real-time data and handle high-dimensional inputs makes it highly suitable for deployment in dynamic urban markets. The development and integration of a user- friendly interface further enhance its practical value for end-users, including real estate professionals, home buyers, and policy-makers.

This study also highlighted key insights such as the importance of data preprocessing, thoughtful feature selection, and the role of model interpretability in ensuring transparency and trust in automated decision-making. The limitations observed—particularly related to data imbalance across geographic and socioeconomic groups—point toward future work in enhancing fairness and generalizability.

In sum, machine learning offers a transformative approach to house price prediction, one that aligns with the demands of data-driven real estate services. As data availability continues to improve, and ethical frameworks evolve, such systems will increasingly become indispensable tools in property assessment and urban development planning.

REFERENCES

[1]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.https://doi.org/10.1145/2939672.29 39785
[2]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/101320 3451
[3]. He, H., & Garcia, E. A. (2009).Learning from imbalanced data. IEEE Transactions on Knowledge and DataEngineering, 21(9), 1263– 1284.
[4]. https://doi.org/10.1109/TKDE.200 8.239
[5]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
[6]. Molnar, C. (2022). Interpretable machine learning: A guide for making black boxmodels explainable (2nd ed.). Leanpub.
[7]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.29 39778
[8]. Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. Proceedings of the 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 1310–1315.
[9]. Tian, Y., & Ma, Y. (2020). House price prediction using machine learning algorithms: A case study of Melbourne housing market. Journal of Big Data, 7(1), 1–16. https://doi.org/10.1186/s40537-020-00326-w
[10]. Zhang, Y., & Wang, J. (2019). Prediction of house prices using ensemble learning models. Procedia Computer Science, 162, 341–347. https://doi.org/10.1016/j.procs.2019.11.288
[11]. Zhu, Y., Lin, T., & Jiang, Y. (2021). Machine learning for housing price prediction: A systematic review. ACM Computing Surveys