House Price Trend Analysis Using Real Estate Data

[1]Mr.M Ramakrishna Raju, [2]B. Hema Lakshmi, [3]A. S Siri Chandana, [4]G. Satya Sindhur, [5]CH. B S Venkat
[1]*Assistance professor, Srinivasa Institute of Engineering and Technology*
[2345] *UG Scholar, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I10-195123-459

Abstract: - This project aims to create a smart way to track and predict house prices using real estate data. In fast-growing cities, knowing the right price is very important for buyers and investors. We found that old-school manual methods of guessing prices often miss how modern home features and market changes work together. To solve this, we used the XGBoost algorithm to make predictions much more accurate. This study utilizes house prices: Advanced Regression Techniques dataset from Kaggle, which contains structured real estate attributes and sale prices. The dataset underwent systematic preprocessing, including handling missing values, categorical encoding, and feature normalization. Exploratory Data Analysis (EDA) was performed to identify key patterns and influential variables affecting property prices. For predictive modeling, the XGBoost regression algorithm was implemented because of its high efficiency, regularization capability, and ability to handle complex nonlinear relationships. To validate the model performance, a comparative analysis was conducted using standard regression metrics. The proposed system achieved a significant Mean Absolute Error (MAE) of approximately $20,532, Mean Squared Error (MSE) of 7.45 X 10^3, and Rsquared(R^2) value of 0.8924, demonstrating high predictive accuracy. Additionally, the system incorporates trend visualization techniques to analyze price variations, enabling a better understanding of market dynamics. This approach provides an accurate, scalable, and practical solution for real estate forecasting, supports informed decision making and reduces uncertainty in property investment.

Key Words: Data Science, Price analysis, Real, Estate, Trends

I.INTRODUCTION

The real estate market serves as a fundamental pillar of the national economy, directly affecting large-scale growth and individual financial security. Determining the value of a property is a complex challenge because house prices are driven by a vast array of interconnected factors ranging from specific attributes such as structural quality and square footage to broader influences such as shifting interest and urban development. These variables create a highly dynamic pricing environment, where market values are constantly evolving. Consequently, developing robust methods to analyze these trends and predict property values accurately has become an essential priority for investors, homeowners, and policymakers alike [1]. Traditionally, property appraisal and house price estimation have relied heavily on manual rule-based methods or simple linear models. While these approaches were sufficient in the past, they often struggle to capture the nonlinear relationships and complex feature dependencies found in modern real estate data [2]. Conventional models frequently fail to account for how various factors such as the synergy between a house's age and its renovated amenities, impact the final market value. Consequently, these methods demonstrate limited predictive accuracy and poor generalization for high-dimensional datasets [3]. To overcome these hurdles, researchers have increasingly turned to advanced machine learning techniques that can learn patterns from historical transaction data. In our research, we implemented machine learning models because they are capable of complex patterns from historical data and are widely applied in regression tasks. Ensemble learning methods have gained attention owing to their ability to improve accuracy by combining multiple models. XGBoost (Extreme Gradient Boosting) is particularly effective for structured datasets and nonlinear feature interactions [4]. By sequentially constructing decision trees and minimizing residual errors, XGBoost achieves high predictive performance while controlling overfitting through advanced regularization techniques [5][7]. This makes it a superior choice for maintaining high accuracy in diverse housing categories. In this study, the XGBoost algorithm was applied to the Ames Housing dataset to analyze price trends and predict residential property

values. This benchmark dataset provides detailed housing attributes, enabling systematic pre-processing and feature engineering to identify key market drivers [6]. Although the model was trained on this data to ensure reliability (achieving an R^2 score of 0.8924), the methodology was designed to be scalable. The objective of this study is to translate theoretical machine learning concepts into a functional framework that can handle real-world market dynamics. By ensuring that the model is adaptable to diverse urban data structures, this study aims to provide a reliable tool for practical real estate forecasting, moving beyond basic academic modelling to offer actionable property value estimations [8].

Recent studies have demonstrated that XGBoost effectively captures complex non-linear relationships in real-world housing data, outperforming traditional regression models [9]. This research also considers broader economic factors and housing indices to improve prediction stability [12].

II. LITERATURE SURVEY

A. Traditional and Hedonic Models

Hedonic pricing models estimate property values based on structural and locational features [6]. Harrison and Rubinfeld [6] showed that these models work well for moderate datasets but lose accuracy with high dimensional urban data, highlighting the need for more approaches.

B. Handling Spares and Complex Data with Random Forests

Modern real estate datasets often contain high dimensional or sparse features that challenge traditional models. Random Forests address this by averaging multiple decision trees, which naturally reduces noise and prevents overfitting [3]. This ensemble approach is highly effective for handling the diverse structural attributes in the Ames housing dataset, where specific architectural details can create outliers. By implicitly performing feature selection, Random Forests provide a robust mechanism for managing complex variables, ensuring stable and reliable property valuations.

C. Gradient Boosting and XGBoost

The Gradient Boosting Machine (GBM) improves predictive power by sequentially minimizing residual errors [5]. XGBoost further optimizes this process by incorporating advanced regularization and effectively captures the nonlinear relationships between a property's physical attributes and its market price [7]. This makes it a highly reliable algorithm for achieving precise valuations without the risk of model overfitting. Further research indicates that optimizing hyperparameters within the XGBoost framework significantly reduce prediction errors compares to Random Forest and other ensemble methods [10].

D. Deep Learning Approaches

While Deep Learning models, such as Multi Layer Perceptron (MLPs), can achieve high accuracy on massive datasets, they often struggle with overfitting when applied to small or medium sized records [5]. For datasets in the 10 K – 20 K range, tree-based models such as Random Forest and XGBoost are generally preferred. These methods provide better interpretability and more consistent performance in structured real estate markets, where volume is limited. While deep learning is powerful tree-based models like XGBoost remain superior for structed datasets, achieving over 90% accuracy in real-time valuation tasks [11].

Research Gap

Most studies have focused on traditional econometrics or general machine learning comparisons. Few studies have implemented a full pipeline using Random Forests or XGBoost for structured datasets such as Ames. This study fills this gap with an XGBoost based framework optimized for complex real estate data.

III. SYSTEM ARCHITECTURE

The proposed architecture is designed as a modular, multi-tier framework to ensure high performance and scalability in property valuation forecasting. The system integrates data management, machine learning logic, and a user-accessible interface into a cohesive pipeline.

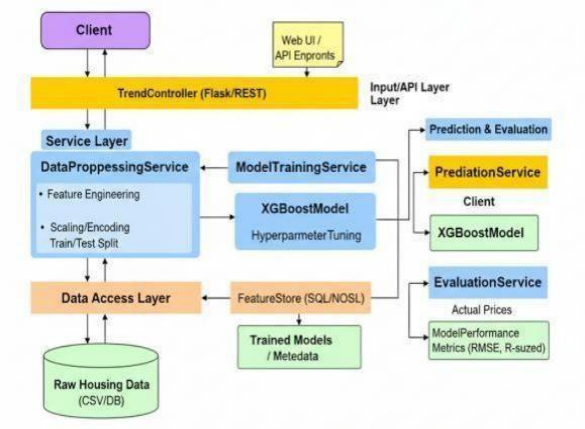

Fig1: System Architecture of XGBoost

1. Data Access Layer

The process begins with Raw Housing Data sourced from the Ames dataset. This layer is responsible for data retrieval and initial storage management of the data. It interacts with a feature store (utilizing SQL / No SQL structures) to manage metadata and maintain a version control over the dataset used for training and validation.

2.Service Layer

This layer acts as the primary engine for the system's intelligence and is divided into two core services:

- Data Preprocessing: This component performs essential cleaning feature engineering and categorical encoding. It also manages the scaling of numerical variables and the train-test split to generalize the model well to unseen data.
- Model Training Service: This service focuses on the XGBoost Model pipeline. It includes a dedicated module for Hyperparameter Tuning, which systematically adjusts the model parameters to minimize the error rates and optimize the predictive stability.

3. Input / API Layer To transition from a theoretical model to a

practical application, the system implements a Trend Controller built on the Flask/REST framework. This layer provides the necessary API endpoint that allows the Web UI to communicate with the back end. This ensures that the inputs are correctly formatted and routed to the prediction engine.

4. Prediction and Evaluation Layer

The final stage of the architecture involves the generation and validation of results.

- Prediction Service: Utilizing the trained XGBoost weights, this service provides real-time house- price estimation based on user- provided features.
- Evaluation Service: To maintain transparency and reliability, this service calculates key Model Performance Metrics, specifically focusing on RMSE calculates key Model Performance Metrics, specifically focusing on RMSE (Root Mean Square Error) and R-squared (R^2), ensuring the system meets established accuracy benchmarks.

IV. METHODOLOGY

We developed our system as an end-to-end machine learning pipeline that transforms raw housing data into reliable predictions. Rather than treating cleaning and model building as separate, manual tasks, we integrated them into an automated workflow that ensures consistency from start to finish.

A. Data Preprocessing and feature Engineering

The first stage of the pipeline uses the Ames Housing dataset, which contains many several details about residential properties. To make sure the data is high quality, the following steps are part of the process:

- Handling Missing Values: The system cleans the data to fill in any gaps so the model training is never interrupted
- Categorical Encoding: Text descriptions of houses are converted into numbers so that the computer can understand them.
- Feature Normalization: All numbers are scaled to a standard to keep the data consistent.
- Exploratory Data Analysis (EDA):

The pipeline looks for patterns and the most important factors that change a house's price.

B. Core Predictive Modeling (XGBoost)

The main part of the predictive pipeline is the XGBoost (Extreme Gradient Boosting) algorithm. This was chosen because it is very fast and prevents the model from making errors.

Ensemble and Sequential Error Reduction Trees: The system works by combining many decision tress to make a very strong and accurate prediction. The Sequential Error Reduction Trees model builds trees one after another, where each new tree fixes the errors

made by the previous tree. It works by combining many small decision tress together to make one very strong and accurate prediction.

Hyperparameter Optimization: To obtain results, the pipeline includes a tuning process that carefully adjustes settings, such as the learning rate and tree depth.

Advanced Regularization: To prevent the model from memorizing the data, XGBoost uses special rules to maintain stable predictions.

Precision Validation: The final part of the pipeline checks the results, achieving a strong Rsquared ($R2$) value of 0.8924 and a mean Absolute Error (MAE) of approximately $20,532.

V. CONCLUSION

In this project, we successfully built a strong machine-learning pipeline to analyze and predict house prices. Using the Ames Housing dataset, we moved away from old manual methods and created a much more accurate, data-driven way to find property values. Using the XGBoost framework worked well for capturing the tricky relationship between house features and their market prices. With an R – Squared (R^2) value of 0.8924 and a Mean Absolute Error of $20,532, our system proved to be very stable and precise. We believe that tool is practical and scalable for real-world applications. It turns complex math into something a regular homeowner or investor can actually use to reduce their financial risk and make better decisions in today's fastmoving markets.

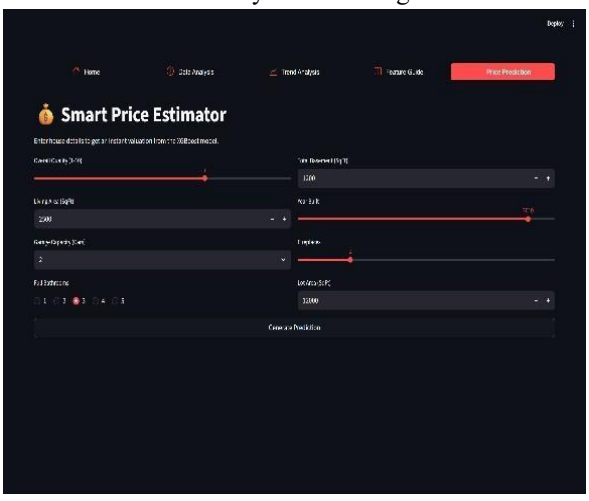

Fig2: Interface Design for Real-Time Feature Acquisition

The figure displays the front-end dashboard where users input property specifications such as living area, quality rating, and amenities. This framework streamlines the data entry process by translating everyday property details into refined numerical inputs, ensuring that the backend model receives clean, standardized data for reliable price estimation.

VI. RESULTS AND DISCUSSION

The performance of the XGBoost pipeline was evaluated using the Ames Housing dataset to ensure its reliability for real-world market forecasting. The system achieved a high level of predictive accuracy, with an R-squared (R^2) value of 0.8924. This indicates that the model can explain approximately 89% of the variance in property prices, demonstrating a strong fit for structured real estate data.

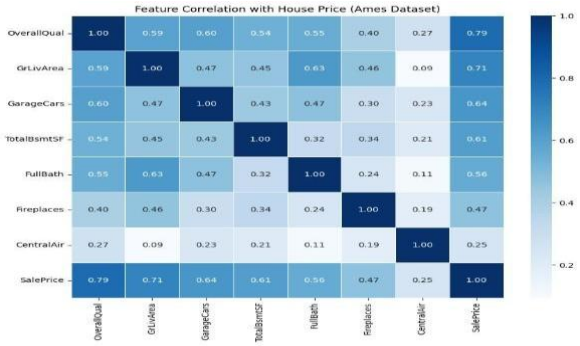

Fig3: Correlation Heatmap of Key Market Drives (Ames Housing dataset)

The model recorded a mean absolute error (MAE) of approximately $20,532 and a Mean Squared Error (MSE) of 7.45 X $10^{3.}$ These metrics confirm that the system maintains a low margin of error making it a dependable tool for estimating property values.

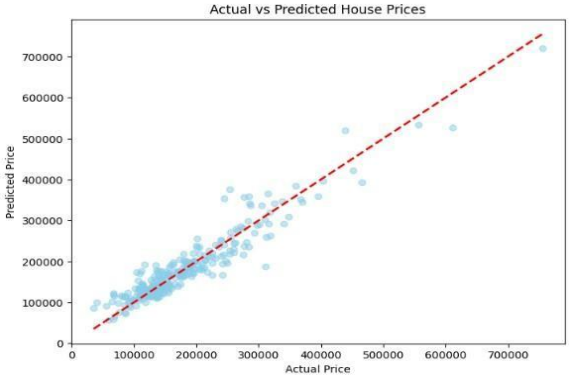

Fig 4: Comparative Analysis of Manual Estimation vs XGBoost Prediction

This graph highlights the gap between traditional manual methods and our advanced machine learning methods and our advanced machine learning methods often struggle with complex feature, the proposed system provides a more stable and accurate valuation. Beyond the numerical results, the findings reveal that the XGBoost algorithm is particularly effective in capturing nonlinear relationships, such as the synergy between a house's age and its modern amenities. By using ensemble learning and sequential error reduction, the system successfully moved beyond basic academic modelling to offer actionable property value estimations.

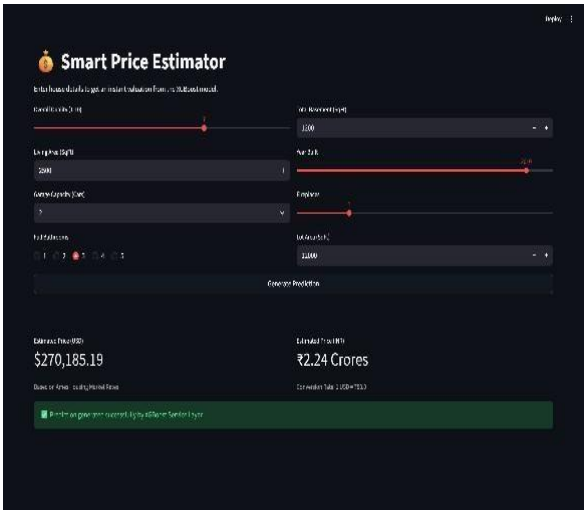

Fig 5: Real-Time Valuation output with MultiCurrency Conversion

The figure displays the successful execution of the predictive pipeline, generating a property valuation of Rs.2.24 Crores ($270,185) based on the input parameters. This output demonstrates the model 's ability to handle high-value property segments and provide actionable financial data for investors and homeowners.

VII. ACKNOWLEDGMENT

We want to give a huge thank you to our project supervisor, R K Raju, for all the technical advice and guidance throughout this research. Their mentorship really helped us turn this predictive framework into a success. We also thank the Department of Computer Science and Engineering for providing the tools and environment needed to complete our work. Finally, we thank our friends and faculty for their valuable feedback while we were building this project.

REFERENCES

[1] J. Friedman, "Multivariate adaptive regression splines," The Annals of Statistics, vol. 19, no. 1, pp. 1–67, 1991.

[2] R. J. Shiller, "Understanding recent trends in house prices and home ownership," Housing, Housing Finance, and Monetary Policy, Federal Reserve Bank of Kansas City, pp. 89– 123, 2007.

[3] S. Mullainathan and J. Spiess, "Machine learning: An applied econometric approach," Journal of Economic Perspectives, vol. 31, no. 2, pp. 87–106, 2017.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189– 1232, 2001.

[6] D. Harrison and D. L. Rubinfeld, "Hedonic prices and the demand for clean air," Journal of Environmental Economics and Management, vol. 5, no. 1, pp. 81–102, 1978.

[7] Kaggle, "Ames Housing Dataset," [Online]. Available at: Kaggle Repository.

[8] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[9] Sandeep Kumar, et al. (2024) "Real Estate House Price Prediction Using Extreme Gradient Boosting" Published in: International Journal of Emerging Technologies and Innovative Research (JETIR).

[10] Hemlata Sharma, et al. (2024) "An Optimal House Price Prediction Algorithm: XGBoost" Published in: Journal of Analytics (MDPI/arXiv).

[11] L.M.I. Leo Joseph, et al. (2024) "Predicting Real-Time House Prices: A Machine Learning Approach Using XGBoost Algorithm" Published in: 2024 Asia Pacific Conference on Innovation in Technology (IEEE).

[12] Singh & Shanmugam (2026/2025) "Exploring predictive models to improve the accuracy of Housing Price Index forecasts in India's real estate sector" Published in: PLOS One.