ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.206

Volume 8, Issue 4, April 2025

© 2025 IJMRSET | Volume 8, Issue 4, April 2025 |DOI:10.15680/IJMRSET.2025.0804472

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Interactive House Price Estimation with XGBoost and Streamlit

Sree Harihara Suthan.K.A, T.R. Anand

Student, Department of Computer Technology, Dr. N. G. P. Arts and Science College, Coimbatore, India

Assistant Professor, Department of Computer Technology, Dr. N. G. P. Arts and Science College, Coimbatore, India

ABSTRACT: The real estate industry is one of the most dynamic and data-rich sectors, where accurate pricing models can significantly benefit buyers, sellers, investors, and financial institutions. Traditional house price estimation methods heavily rely on human judgment, which can lead to inconsistent, biased, and inaccurate valuations. This paper presents a House Price Prediction System developed using Python that utilizes advanced machine learning techniques, specifically the XGBoost Regressor, to deliver highly accurate property price estimations. The system incorporates a structured pipeline that includes data preprocessing, missing value handling, categorical encoding, and feature selection using correlation analysis to improve model reliability and performance.The model is trained and validated on a real-world housing dataset, and its performance is evaluated using key metrics such as Mean Absolute Error (MAE) and the R^2 Score. To enhance accessibility and usability, the trained model is deployed using Streamlit, creating an interactive web-based application that enables users to input property attributes—such as lot area, number of rooms, garage space, and construction quality—and receive real-time price predictions. The application also provides visual insights through correlation heatmaps, feature importance graphs, and missing value charts, enabling users to understand the underlying factors influencing house prices.This system demonstrates how machine learning can revolutionize the real estate valuation process by offering scalable, data-driven, and user-friendly tools. The approach not only ensures better transparency and efficiency but also opens avenues for integrating live market data, location intelligence, and predictive analytics for future development. The results affirm that the proposed model is highly capable of addressing practical real estate challenges, positioning it as a valuable asset for real-time decision-making and strategic planning in property investment.

Keywords: Machine Leraning,XGBoost,Streamlit,Data Preprocessing,Real Estate,Regresssion.

I. INTRODUCTION

The valuation of real estate assets is a critical task in urban planning, property investment, and individual financial decision-making. With urbanization and digitization, there has been a surge in the availability of real estate datasets that include diverse property attributes such as land area, building quality, number of bedrooms, garage space, and neighborhood details. These datasets provide an opportunity to apply machine learning (ML) techniques to automate and optimize the prediction of house prices. However, several challenges persist. Real estate data is often incomplete, high-dimensional, and heterogeneous**,** containing both numerical and categorical attributes.

Manual analysis of such data is not only labor-intensive but also susceptible to human bias and error. The need for a scalable, automated, and intelligent system that can learn from past trends and predict property prices accurately is more pressing than ever, especially in rapidly growing economies. Machine learning provides a significant advancement over traditional valuation approaches. Unlike rule-based systems, ML models can learn complex patterns from historical data, identify non-linear relationships, and generalize these learnings to new, unseen data. Among the many ML models, XGBoost has emerged as a top performer due to its robust architecture, which incorporates gradient boosting along with regularization and tree pruning, resulting in better accuracy and generalization.

© 2025 IJMRSET | Volume 8, Issue 4, April 2025 |DOI:10.15680/IJMRSET.2025.0804472

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

In this study, we propose a House Price Prediction System using Python, which not only achieves high prediction accuracy using XGBoost but also emphasizes usability by deploying the model using Streamlit, a modern tool for building data-driven web apps. This dual focus on model performance and accessibility makes the system practical for widespread use. Furthermore, the system integrates data visualization modules, offering insights into which property features most influence the price, thereby enhancing transparency and interpretability. Such features are especially important for stakeholders who may not have a technical background but require decision support in real estate investments.

II. LITERATURE REVIEW

The application of machine learning in real estate has evolved considerably in the last decade. Earlier studies predominantly focused on hedonic pricing models and multiple linear regression**,** which assume independence between features and linear relationships with the target variable. While useful in certain contexts, these models fall short in capturing the complexity of modern housing markets.

Recent works have turned to tree-based algorithms for their ability to model non-linear relationships, handle missing values, and manage both categorical and continuous data. For example, Random Forests, as used in the study by Kumar et al. (2021), showed improvements in predictive accuracy over linear regression by aggregating multiple decision trees. However, they still suffered from interpretability issues and overfitting in some scenarios.

To mitigate these limitations, researchers have increasingly adopted boosting algorithms**,** particularly XGBoost. Introduced by Chen and Guestrin (2016), XGBoost implements advanced techniques like second-order optimization**,** column block structure for parallel computation**,** and tree pruning, which make it both efficient and highly accurate**.** A comparative study by Singh and Srivastava (2023) showed that XGBoost consistently outperformed Gradient Boosting Machines (GBM), AdaBoost, and Support Vector Machines (SVM) in predicting housing prices.

The use of feature engineering and selection is another critical theme in recent literature. Several papers highlight the need to preprocess real estate data by removing irrelevant or redundant features, transforming categorical attributes into numerical ones, and using correlation analysis to identify impactful predictors. Feature importance rankings derived from ensemble models further support this process by quantifying the relative impact of each feature on the target variable.
An emerging focus in machine learning is the deployment and usability of models. While many studies report high accuracy, they often overlook accessibility for end users. Streamlit, introduced in 2019, addresses this gap by enabling rapid deployment of ML models as interactive web applications without requiring advanced front-end development skills. This broadens accessibility, allowing real-time predictions for non-technical users in fields such as real estate.

Equally important is model interpretability. Despite the predictive power of models like XGBoost, their complexity can hinder understanding. Visual tools such as correlation heatmaps, feature importance graphs, and SHAP value plots are increasingly used to explain predictions. For instance, Zhang et al. (2022) leveraged SHAP values to clarify the impact of specific features on price predictions, enhancing user trust and transparency.

In summary, the literature clearly supports the transition from traditional, manual methods to automated, intelligent, and interpretable machine learning systems**.** However, many of these implementations are not readily usable by real estate practitioners or end-users. This paper addresses this critical gap by combining the predictive strength of XGBoost with a user-centric web interface**,** thereby offering a practical and effective solution for modern house price prediction.

III. METHODODLOGY

The proposed House Price Prediction System is designed as a full-cycle machine learning application, combining advanced data preprocessing, predictive modeling, and interactive deployment. The system leverages the capabilities of the XGBoost Regressor for accurate predictions and utilizes Streamlit for a user-friendly web-based interface. The entire architecture is modular and scalable, making it suitable for practical use in real estate applications.

© 2025 IJMRSET | Volume 8, Issue 4, April 2025 |DOI:10.15680/IJMRSET.2025.0804472

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

1) Data Collection

A structured dataset is acquired from an open-source platform such as Kaggle, containing detailed records of residential property attributes. These include both numerical features (e.g., lot area, garage capacity) and categorical features (e.g., neighborhood, building type). The dataset serves as the foundation for model training and evaluation.

2) Data Preprocessing

Preprocessing is a critical step to ensure data quality and model performance. This phase includes:

- **Missing Value Handling:** Columns with excessive missing data are removed, while columns with minimal missing data are imputed using statistical methods such as mean substitution.
- **Categorical Encoding:** Non-numeric data is transformed into numeric format using Label Encoding, enabling compatibility with the XGBoost model.
- **Feature Selection:** Highly correlated features are retained based on correlation matrix analysis to enhance model performance and reduce dimensionality.

3) Model Training – XGBoost Regressor

The preprocessed data is divided into training and testing subsets (typically 80:20). The XGBoost Regressor, known for its regularization capabilities and boosting performance, is trained to predict the Sales Price target variable. Hyperparameters such as learning rate, tree depth, and estimators are optimized to reduce bias and variance.

4) Model Evaluation

The trained model is evaluated using the following metrics:

- **Mean Absolute Error (MAE):** Indicates the average absolute difference between predicted and actual values.
- **R^2 Score:** Reflects how well the model explains the variability in house prices.

A strong R^2 value (e.g., 0.89) and a low MAE confirm that the model is reliable and ready for deployment.

5) Model Saving

The optimized model is serialized using the joblib library. This allows for fast loading during inference without the need for retraining, making the application more responsive and resource-efficient.

6) Streamlit Web Interface Deployment

To enable real-time interaction with end-users, the model is deployed using Streamlit, a Python-based web app framework. Users can enter property features through an intuitive web form. The model processes the inputs, predicts the house price, and displays the result along with:

- Feature Importance Graphs
- Correlation Heatmaps
- Prediction Output

This deployment transforms the model into a fully functional application, accessible to real estate agents, buyers, sellers, and analysts without the need for programming knowledge.

IV. SYSTEM ARCHITECTURE

LEVEL 0 : TRAINING MODEL

The Level 0 Data Flow Diagram (DFD) gives a simple overview of the House Price Prediction System. It starts with Data Collection & Preprocessing, where raw data is cleaned, missing values are handled, and unnecessary columns are removed. Next, Correlation Analysis selects the most important features that affect house prices. The refined data is then used in

© 2025 IJMRSET | Volume 8, Issue 4, April 2025 |DOI:10.15680/IJMRSET.2025.0804472

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Model Training, where the XGBoost Regressor learns from past house prices.After training, the model goes through Prediction & Evaluation, where its accuracy is tested using Mean Absolute Error (MAE) and R^2 Score. Finally, the trained model is saved using Joblib so it can be used in a web app for real-time price predictions.

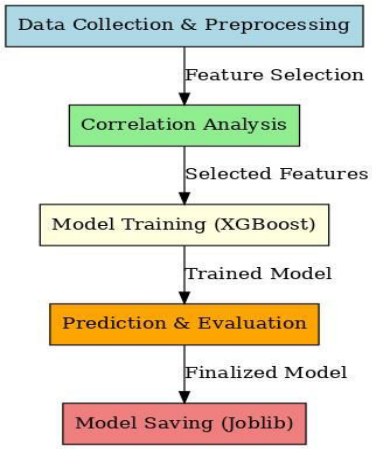

Figure 1: Model Training Work Flow

LEVEL 1 : CREATING WEB INTERFACE

The Level 1 Data Flow Diagram (DFD) provides a detailed view of how the House Price Prediction System operates within the web application.

The process starts with Loading the Trained Model, where the saved XGBoost model (using Joblib) is loaded to make predictions. Next, the Web Interface is Built with Streamlit, providing a user-friendly platform where users can enter property details.

When a user inputs data, the system moves to User Input Handling, where the input values are processed and formatted correctly before being sent to the model. The model then makes a prediction in the Prediction & Display phase, where the estimated house price is calculated and shown on the web interface.Finally, the system runs smoothly in the Running the Web App stage, ensuring real-time predictions and seamless user experience.

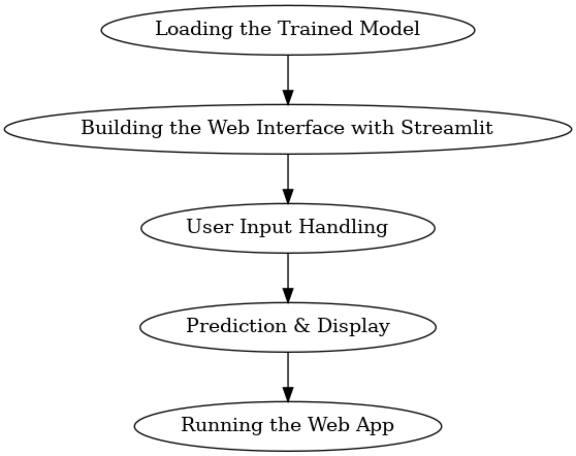

Figure 2: creating web interface

© 2025 IJMRSET | Volume 8, Issue 4, April 2025 |DOI:10.15680/IJMRSET.2025.0804472

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. DATA FLOW DIAGRAM

The Data Flow Diagram (DFD) outlines the workflow of the House Price Prediction System. Users input property details via a Streamlit interface, which are then preprocessed—handling missing values, encoding categories, and selecting features—to prepare the data for modeling.The cleaned data is fed into an XGBoost Regressor to generate price predictions. Model performance is evaluated using MAE and R^2 Score, and results are visualized with tools like feature importance plots and correlation heatmaps, making outputs both accurate and interpretable.This system not only streamlines real estate valuation but also enhances user accessibility and trust. By combining robust machine learning techniques with a simple web interface and clear visual explanations, it empowers users—including buyers, sellers, and agents—to make informed, data-driven decisions in the property market.

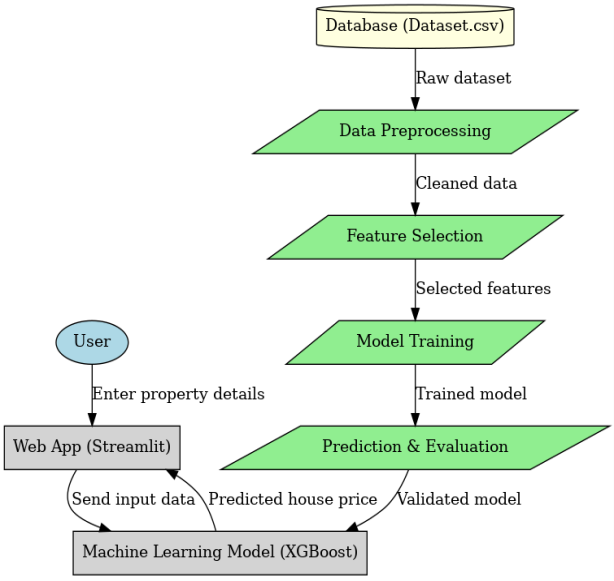

Figure 3: Workflow of the interface

VI. IMPLEMENTATION

The House Price Prediction System was implemented using Python and various open-source libraries. The core machine learning model was built using the XGBoost Regressor, chosen for its accuracy and efficiency in regression tasks.

The dataset was sourced from Kaggle, containing historical house data with features such as LotArea, OverallQual, TotalBsmtSF, and GarageArea. Initial implementation steps included data preprocessing, where irrelevant columns (e.g., Id, Alley, PoolQC) were dropped, and missing values (e.g., LotFrontage) were handled using mean imputation. Categorical variables were encoded using Label Encoding for compatibility with the model.

Feature selection was performed using correlation analysis, retaining only the attributes highly correlated with the target variable (SalePrice). A correlation heatmap was plotted to visualize the relationships among features, and a feature importance chart was generated after model training to identify the most influential attributes.

The data was split into training and testing sets in an 80:20 ratio. The XGBoost model was trained and evaluated using Mean Absolute Error (MAE) and R^2 Score to assess its prediction capability. The trained model was saved using the Joblib library for reuse.

© 2025 IJMRSET | Volume 8, Issue 4, April 2025 |DOI:10.15680/IJMRSET.2025.0804472

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

For end-user interaction, a simple and interactive web interface was developed using Streamlit. Users can enter property details through sliders and input fields, and the app displays real-time price predictions. The app also includes visual feedback such as prediction results and feature analysis, ensuring the system is both informative and user-friendly.

VII. UNIQUNESS

The House Price Prediction System developed in this study offers several advantages over traditional and existing prediction methods:

1. Enhanced Accuracy with XGBoost: Unlike simple linear regression models, this system uses XGBoost, a powerful gradient boosting algorithm that handles non-linearity and feature interactions more effectively, leading to improved prediction accuracy.

2. Smart Data Preprocessing: The system applies a structured data cleaning pipeline—handling missing values, encoding categorical variables, and performing feature selection based on correlation analysis—which ensures that only meaningful and clean data is used for model training.

3. Interactive and Real-Time Predictions: The use of Streamlit allows users to interact with the model in real-time, making the system accessible and user-friendly for individuals with no technical background. Most existing systems lack this interactive frontend.

4. Visual Insights for Transparency: By integrating visual tools like correlation heatmaps and feature importance plots, users can understand which factors influence house prices, adding a layer of interpretability often missing in other models.

5. Lightweight and Cost-Effective: The entire system runs on minimal hardware and does not require heavy infrastructure or paid cloud services. This makes it ideal for educational use or small-scale deployment.

6. Scalable for Future Enhancement: The modular architecture of the project allows easy integration of additional features like live market data, map-based location inputs, or more advanced ensemble models in the future.

VIII. CONCLUSION

This project successfully presents a House Price Prediction System that applies advanced machine learning techniques to provide accurate and efficient real estate valuation. By using the XGBoost Regressor, the model captures non-linear relationships between diverse property attributes, outperforming traditional linear models in prediction accuracy.

A critical part of the system's success lies in its structured data preprocessing pipeline. This includes handling missing values, encoding categorical variables, and selecting highly correlated features. These steps ensure that the input data is clean, relevant, and optimized for the machine learning model, contributing significantly to performance.To make the model accessible to a broader audience, a Streamlit-based web interface was developed. This interface allows users—regardless of their technical background—to input property details and receive real-time price predictions. The inclusion of visual tools such as correlation heatmaps and feature importance plots also improves interpretability and user trust.

The system is designed to be lightweight and cost-effective, requiring minimal resources for deployment. Its modular structure makes it easy to maintain and extend, allowing for the integration of additional features like live real estate market data, geospatial inputs, or advanced ensemble models in future versions.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

In summary, the proposed system bridges the gap between machine learning and practical usability in the real estate domain. It provides a scalable, interpretable, and user-friendly solution that supports data-driven decision-making for buyers, sellers, and property analysts, marking a significant step toward smarter and more transparent property valuation.

REFERENCES

1. Sharma, H., Harsora, H., & Ogunleye, B. (2024). An Optimal House Price Prediction Algorithm: XGBoost. Analytics, MDPI. Available at: https://www.mdpi.com/2813-2203/3/1/3
2. Wang, S., Ma, Y., & Zhang, Z. (2020). Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science, 174, 1914–1923. Available at: https://www.sciencedirect.com/science/article/pii/S1877050920316318
3. Abdul-Rahman, S., & Mutalib, S. (2021). Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur. International Journal of Advanced Computer Science and Applications (IJACSA), 12(12), 204–210. Availableat:https://www.researchgate.net/publication/357455077_Advanced_Machine_Learning_Algorithms_for_House_Price_Prediction_Case_Study_in_Kuala_Lumpur
4. XGBoost Documentation – Implementation of XGBoost Regressor. Available at: https://xgboost.readthedocs.io/
5. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD, 785–794. https://xgboost.readthedocs.io/
6. Hasan, M. H., Jahan, M. A., Ali, M. E., Li, Y.-F., & Sellis, T. (2024). A Multi-Modal Deep Learning Based Approach for House Price Prediction. arXiv preprint. Available at: https://arxiv.org/abs/2409.05335

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com