International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 04 | Apr 2024 www.irjet.net p-ISSN: 2395-0072

# House Price Prediction Based On Machine Learning

**Chandan Barik, Deepak Kumar Sahu, Rajat Dewangan, Priyanka Devi**

*B.Tech Student, Dept. Of Information Technology, Govt. Engineering College, Bilaspur, CG., India*
*B.Tech Student, Dept. Of Information Technology, Govt. Engineering College, Bilaspur, CG., India*
*B.Tech Student, Dept. Of Information Technology, Govt. Engineering College, Bilaspur, CG., India*
*Assistant Professor, Dept. Of Information Technology, Govt. Engineering College, Bilaspur, CG., India*

---

**Abstract -** *This study explores the utility of house price prediction in facilitating informed decision-making for both developers and potential buyers. While the House Price Index (HPI) is a widely employed tool for estimating housing price fluctuations, the intricate relationship between housing prices and various factors such as location, size, and demographics necessitates additional data for accurate individual price predictions. Despite the abundance of research utilizing traditional machine learning methods to forecast housing prices, there is a notable lack of focus on evaluating the performance of individual models and a tendency to overlook more sophisticated, albeit less mainstream, approaches.*

*The Global House Price Prediction System project aims to enhance efficiency, accuracy, consistency, and risk mitigation in decision-making processes related to house price predictions. Key objectives include streamlining processes, leveraging historical data and machine learning for accuracy, ensuring consistency in decision-making, minimizing lending risks, creating a user-friendly interface, implementing robust security measures, and conducting thorough testing and validation.*

***Key Words***: Correlation Analysis, Mitigation, Regression, Scalability, Supervised Learning, Outliers, Price Index.

## 1. INTRODUCTION

Typically, the House Price Index captures the aggregated fluctuations in residential property values. To enhance the ease of house hunting for families, we have refined the process by soliciting specific criteria such as desired square footage, number of bedrooms, and bathrooms. Employing preloaded datasets and innovative data features, this paper explores practical data preprocessing and inventive feature engineering techniques. Additionally, it introduces a regression technique within machine learning to forecast house prices. The primary advantages of this project lie in addressing the conservative budgeting and market strategies of prospective homebuyers. It aims to streamline operations and enhance efficiency, offering customers a swift and reliable method for determining house prices. By ensuring transparency and fairness, the project seeks to prevent users from being misled or exploited. Ultimately, its goal is to empower users to search for homes within their budgetary constraints, facilitating a smoother and more informed homebuying process.

### 1.1 Navigating the Complexity of Market Dynamics

Real estate markets are subject to intricate webs of economic, social, and environmental forces. Crafting a model capable of comprehensively capturing and analyzing these dynamics poses a formidable challenge.

### 1.2 Data Variability and Quality

Global house price data may originate from a myriad of sources, each varying in reliability and completeness. Tackling the task of cleaning, preprocessing, and harmonizing data across diverse regions and formats presents a substantial hurdle.

### 1.3 Temporal and Spatial Dynamics

Housing markets exhibit temporal and spatial fluctuations, with distinct regions experiencing unique trends and cycles. A robust model must adeptly accommodate these variations to furnish precise predictions across heterogeneous geographical landscapes

### 1.4 Feature Selection and Engineering

The discernment of pertinent features and their seamless integration into the model is pivotal for ensuring prediction accuracy. Feature engineering plays a pivotal role in transforming raw data into meaningful predictors capable of capturing underlying patterns.

### 1.5 Model Generalization

Crafting a model that not only excels on historical data but also extrapolates effectively to unseen data is paramount. Striking a balance to mitigate overfitting or underfitting, which could lead to erroneous predictions, represents a formidable challenge

### 1.6 Interpretability and Explainability

Fostering transparency in the predictive process holds paramount importance, particularly in the realm of real estate where decisions carry significant ramifications. The

development of an interpretable model capable of elucidating the factors influencing predictions emerges as a pivotal endeavor.

### 1.7 Scalability

Ensuring the model's scalability to accommodate vast swathes of data from heterogeneous sources is imperative. This scalability not only enhances applicability to global markets but also facilitates adaptability to evolving data landscapes over time.

## 2. METHOD

2.1 Logistic Regression:-Logistic regression, a cornerstone of supervised learning, excels in classification endeavors, tasked with predicting the likelihood of an instance belonging to a specific class. Leveraging a sigmoid function on the linear regression output, it aptly earns its name despite its classification focus. The process of logistic regression modeling involves several key steps:

Problem Definition: Begin by identifying the dependent and independent variables, determining whether the problem is binary classification.

Data Preparation: Clean and preprocess the data to ensure it's suitable for logistic regression analysis.

Exploratory Data Analysis (EDA): Visualize relationships between variables, detecting outliers or anomalies.

Feature Selection**:** Choose independent variables with significant relationships to the dependent variable, eliminating redundant or irrelevant features.

Model Building: Train the logistic regression model on the selected variables, estimating coefficients.

Model Evaluation: Assess model performance using metrics like accuracy, precision, recall, F1-score, or AUC-ROC.

Model Improvement: Fine-tune the model based on evaluation results, adjusting variables, adding new features, or employing regularization to counter overfitting.

Model Deployment: Deploy the logistic regression model in practical scenarios, making predictions on new data. The accuracy of this model is 80%.

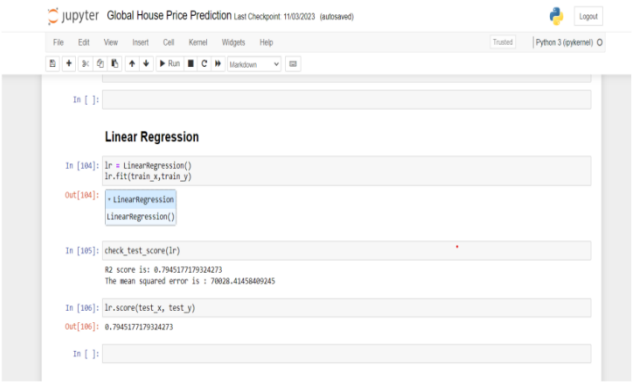



**Chart -1**: Linear Regression

2.2 Correlation Analysis:-Correlation analysis is a statistical technique used to evaluate the strength and direction of the relationship between two quantitative variables. It helps identify whether and how much two variables change together. The correlation coefficient, usually denoted by

r, ranges from -1 to 1, where:

=

1

r=1: Perfect positive correlation (as one variable increases, the other also increases linearly).

=

–

1

r=−1: Perfect negative correlation (as one variable increases, the other decreases linearly).

=

0

r=0: No correlation (the variables do not have a linear relationship).

Correlation analysis is crucial in various fields such as economics, finance, psychology, and epidemiology, among others, to understand relationships between variables and make informed decisions.


**Chart -2**: Correlation Analysis

2.3 Random Forest Algorithm:- Random Forest stands as a widely embraced machine learning algorithm within the realm of supervised learning. Its versatility extends to both Classification and Regression tasks in ML. Rooted in the principle of ensemble learning, it harnesses the power of multiple classifiers to tackle intricate problems and elevate model performance. The accuracy of this model is 98%.

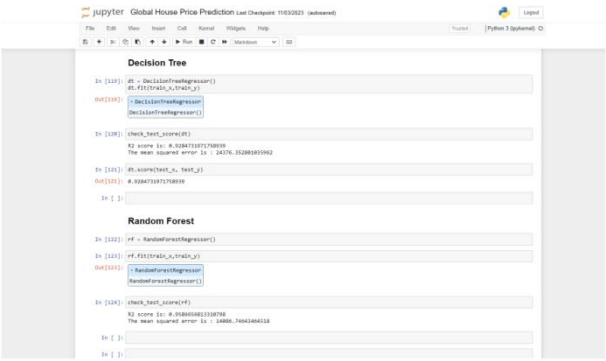



**Chart -3**: Random Forest & Decision Tree

2.4 Decision Tree:- A cornerstone of supervised learning, the decision tree method proves invaluable for both classification and regression endeavors. Constructing a tree-like structure, it employs internal nodes for attribute tests, branches for outcomes, and terminal nodes for class labels. Through recursive partitioning of training data based on attribute values, it iteratively refines its structure until meeting predefined criteria like maximum tree depth or minimum node samples. Analyzing datasets begins at the root node, assessing attribute values against dataset records, and navigating branches until reaching leaf nodes. This iterative process continues through subsequent nodes, comparing attributes with sub-node values until arriving at the tree's terminus. The accuracy of this model is 93%.

## 3. RESULTS & PERFORMANCE ANALYSIS

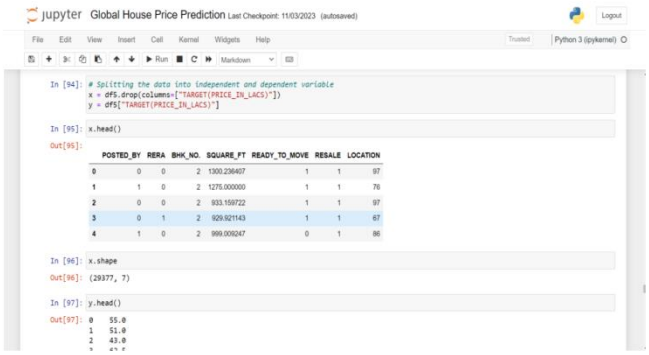



**Fig -1**: Train & Test Data

The training data is used to train the model by feeding it with examples and their corresponding correct outcomes, allowing the model to learn patterns and relationships within the data. This phase involves adjusting the model's parameters to minimize errors and optimize its performance.

On the other hand, the test data also referred to as the validation or evaluation set, is kept separate from the training data and is used to assess the model's performance on unseen data. By evaluating the model's predictions against the actual outcomes in the test set, researchers can gauge its generalization ability and how well it performs on new, unseen data

"Data Visualization" is the graphical representation of data and information to communicate insights and patterns effectively. It serves as a powerful tool for understanding complex datasets, identifying trends, and making data-driven decisions. data visualization is a multifaceted process that involves understanding the data, selecting appropriate visualization techniques, applying design principles, incorporating interactivity, and storytelling, considering the audience, and leveraging relevant tools and technologies to effectively communicate insights and facilitate data-driven decision-making.

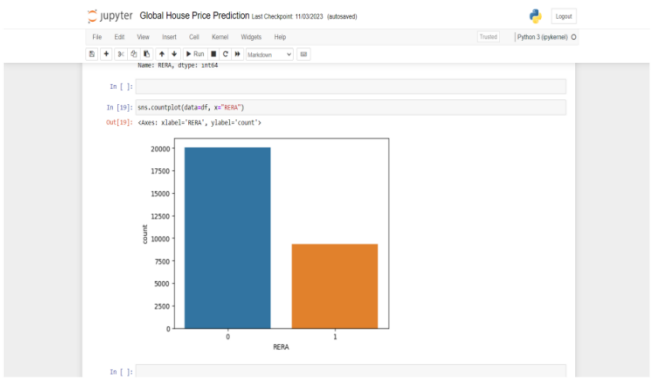



**Fig -2**: Data Visualization

Model Building Using Gradient Boosting" refers to the process of constructing predictive models using the gradient boosting technique, a powerful machine learning algorithm known for its effectiveness in various tasks, including regression and classification.

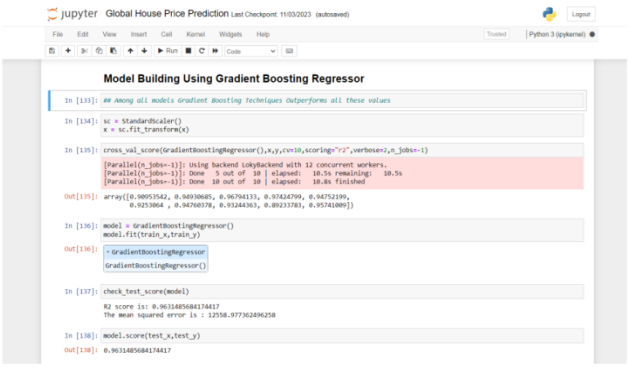



**Fig -3**: Model Building Using Gradient Boosting

Gradient boosting is an ensemble learning method that works by combining multiple weak learners, typically decision trees, to create a strong predictive model. The process involves sequentially adding new models to correct the errors made by previous ones. Each new model is trained on the residuals (the differences between the predicted and actual values) of the previous models, thereby gradually reducing the overall prediction error.

The "web application of house price prediction model" abstract encapsulates the development and deployment of an online platform where users can input relevant data about a property, and the application provides a prediction of the property's price based on a machine learning model.

Data Input Interface: The web application provides a user-friendly interface where users can input information about the property they are interested in, such as location, size, number of bedrooms, amenities, and other relevant features.

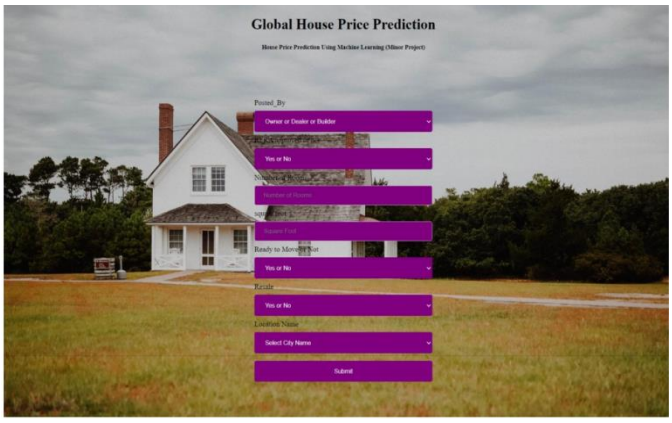



**Fig -4**: Web Application of The Model

This interface includes dropdown menus, text fields, sliders, or interactive maps for selecting locations.

Integration with Machine Learning Model: Behind the scenes, the web application integrates with a machine learning model trained on historical housing data. This model uses algorithms such as regression or gradient boosting to analyze the input data and predict the price of the property.

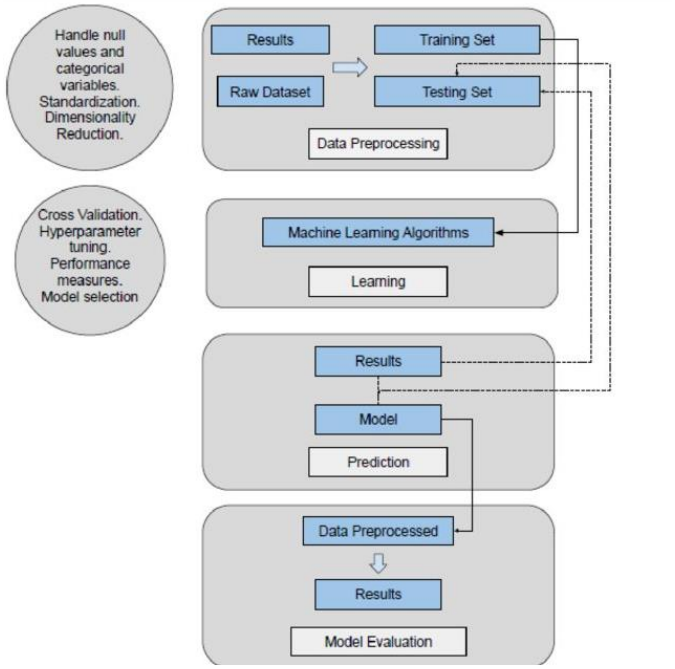


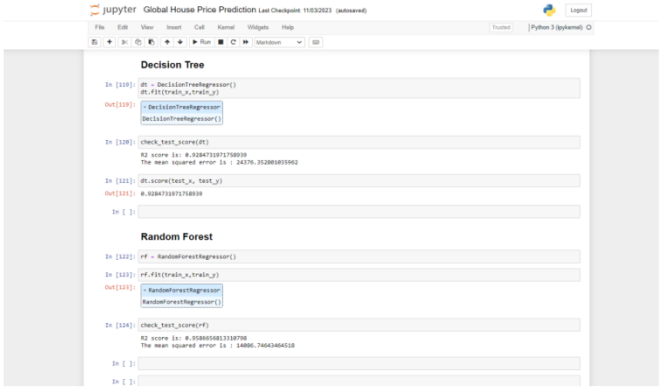

**Fig -5**: System Architecture

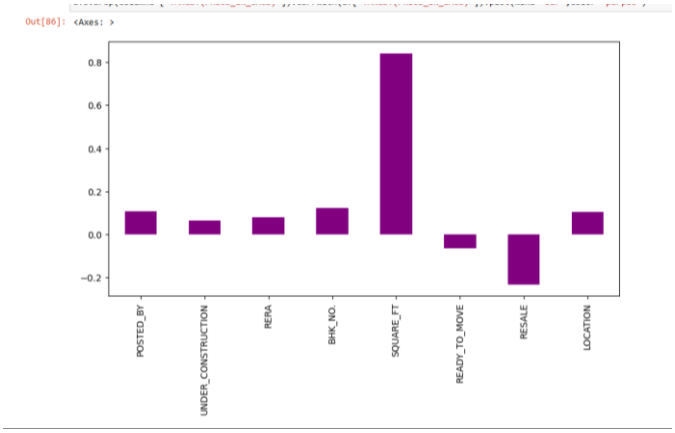
Algorithms and Their Performances:

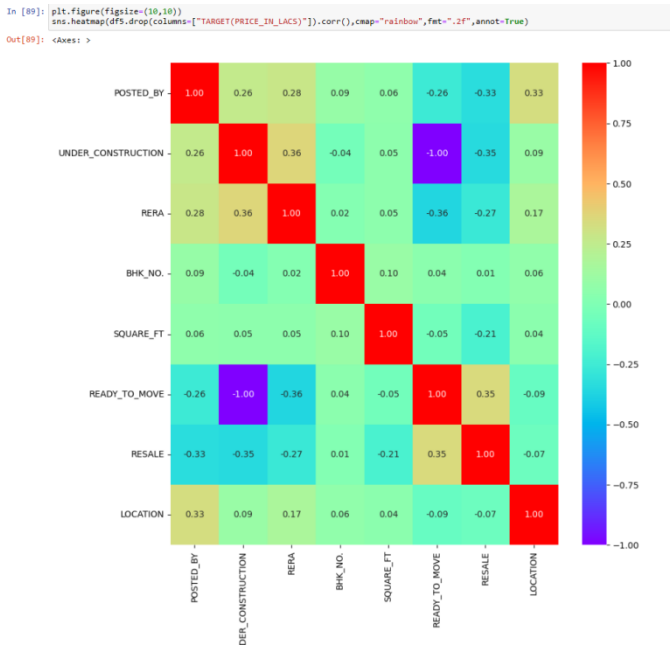
Linear Regression – 80% Accuracy

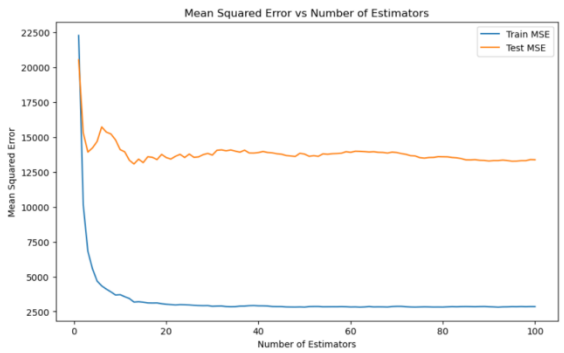
Decision Tree – 93% Accuracy

Random Forest – 96% Accuracy






## 4. CONCLUSIONS

A worldwide initiative focused on forecasting housing prices could prove immensely beneficial for individuals, investors, and stakeholders within the real estate sector, offering valuable insights into property valuations, market trends, and risk assessment. Nonetheless, the effectiveness and significance of this endeavor are contingent upon several crucial factors.

Data Quality.

- Machine Learning Models.
- User-Friendly Interface.
- Transparency and Trust.

Overall, these potential developments reflect the evolving nature of the real estate market and advancements in technology, indicating a promising future for global house price prediction projects. These enhancements aim to improve the accuracy, reliability, and usability of prediction models while leveraging new data sources and analytical techniques to address the complexities of the real estate market.

## REFERENCES

[1] House Price Index. Federal Housing Finance Agency. https://www.fhfa.gov/ (accessed September 1, 2019)M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Fan C, Cui Z, Zhong X. House Prices Prediction with Machine Learning Algorithms. Proceedings of the 2018 10th International Conference on Machine Learning and Computing ICMLC 2018. doi:10.1145/3195106.3195133

[3] Phan TD. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 2018 International Conference on Machine Learning and Data Engineering (ICMLDE) 2018. doi:10.1109/icmlde.2018.00017.

[4] Mu J, Wu F, Zhang A Housing Value Forecasting Based on Machine Learning Methods Abstract and Applied Analysis,

[5] Wolpert D H Stacked generalization Neural Networks, 5 (1992), pp. 241-259

[6] https://profile/ValderiLeithardt/publication/343500604/figure/fig1/AS:921812657520642@1596788672330/Roadmap-for-applying-Machine-Learning-algorithms-in-predictive-analysis-Raschka-and.png