IJARSCT
ISSN: 2581-9429
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

Machine Learning Approach for House Price Prediction

Mrs . Pooja Chaudhary, Dr. Preeti Sharma, Ashutosh, Harsh Chauhan, Abhay Pundir, Himanshu
CSE (Data Science)
Raj Kumar Goel Institute of Technology, Ghaziabad, India

Abstract: *The real estate industry is unique, and predicting house prices is a key part of it. Being able to pull information from raw data makes it much easier to predict prices and understand what makes a house valuable. However, house prices can change daily, sometimes going up unexpectedly. These changes can significantly affect homeowners and the real estate market. The research suggests that artificial neural networks, support vector regression, and linear regression are the most effective ways to model prices. It also indicates that real estate agents and location play a big role in determining how much a property costs. This study can help housing developers and researchers by identifying the most critical factors in housing prices and the best machine learning model to use.*

Keywords: real estate industry

I. INTRODUCTION

Finding a place to live is a basic human need, right up there with food and water. In the real estate world, being able to predict how much a house is worth is super important. It helps both buyers and sellers make smart decisions. Thanks to advancements in machine learning, we've developed many algorithms to accurately estimate property prices. In this study, we're using a dataset of real estate properties along with XGBoost, a really powerful gradient boosting technique, to predict house values [3-5]. XGBoost is great at handling complex datasets [6-8]. It's known for doing well in forecasting and has been used in many machine learning competitions. In our experiment, we're using XGBoost to tackle the challenge of predicting housing prices and see how well it performs.

The goal of predicting house prices is to create a model that can accurately estimate the price of a new house. It does this by looking at previous data on house features (like square footage, number of bedrooms and bathrooms, location, etc.) and their prices. In this project, we've applied five algorithms: linear regression, support vector machine, Lasso regression, Random Forest, and XGBoost**.**

1.1 Explanation

Input: The input section represents the initial stage of the house price prediction process. Here, one has to gather relevant data that could influence house prices, it can be the dataset. This data may include factors such as the size of the house, number of bedrooms, location, neighborhood amenities, historical sales data, and other relevant features.

Preprocessing: In the preprocessing stage, the collected data goes through various cleaning and transformation steps to ensure its quality and suitability for analysis. In this stage several tasks perform like handling missing values, removing outliers, normalization or scaling the data and encoding categorical variables. Preprocessing helps to prepare the data for effective modelling.

DOI: 10.48175/IJARSCT-26994

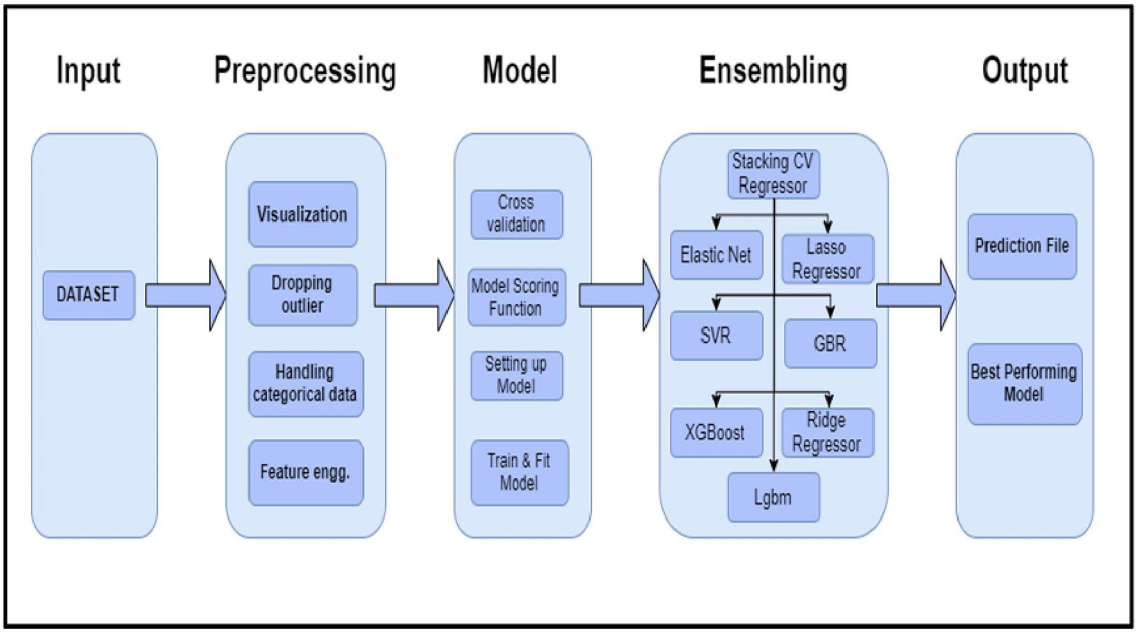


Fig. 1. Flow of execution

Model: The model section represents the core of the house price prediction process. Here, one has to select an appropriate machine learning algorithm or ensemble of algorithms to build a predictive model. We pick a machine learning algorithm to build that can estimate prices. Some popular algorithms include linear regression, decision trees, random forests, support vector machines, and neural networks. This model takes the prepared data as input and learns from it. It identifies patterns and relationships in the data to predict house prices.

Ensembling: Ensembling refers to combining multiple models to make even better predictions. By using different models together, we can reduce errors and get more reliable results.

Output: The output section represents the final stage of the house price prediction process. Here, the trained model or ensemble provides predictions on house prices based on the given input data. The predictions can be in the form of specific price values or in percentage.

II. PROBLEM STATEMENT

The challenges in accurately predicting house prices stems from the fact that the asking price and general property descriptions are often presented separately from standardized real estate attributes. This disconnect makes it difficult to create a clear and consistent picture of a property's true value, hindering effective price prediction.

III. SYSTEM DESIGN AND ARCHITEC TURE

3.1 Phase I: Collection of Data

In this phase, relevant data is gathered from the different sources such as real estate websites and public datasets. The data may include parameters such as location, number of rooms, number of bathrooms, area and sale prices. The data should be diverse and representative of the target market.

3.2 Phase II: Data Pre-processing

In this phase, the collected data is cleaning and preparing for the model training. There are several task perform in this phase such as handling missing values, removing outliers, normalization or scaling the numerical features, and encoding categorical variables are performed. Feature selection techniques are also applied to identify the most relevant attributes for predict the house prices. Additionally, data splitting techniques such as stratified sampling can be used to create training and testing datasets.

3.3 Phase III: Training the Model

In this phase, different types of machine learning algorithms are applied to train this model using the cleaned data. Typical methods include linear regression, decision trees, random forests, and more sophisticated techniques like gradient boosting or neural networks. The training process encompasses adjusting the model to the training data, fine-tuning hyperparameters, and assessing how well the model performs using relevant metrics like mean squared error or R-squared. The training set is used to help the machine learning model to learn the connections and patterns between the input features such as the number of rooms, location, square footage, and the target variable.

3.4 Phase IV: Testing the Model

Once the model has been trained, it's time to evaluate it using the testing dataset to understand its predictive abilities. We compare what the model predicts with the actual house prices from the testing set to gauge its performance. To determine how accurate the predictions are, we can use evaluation metrics like mean absolute error or root mean squared error.

The testing set is a distinct portion of the dataset, specifically set aside to measure how well the trained model performs and how it generalizes to new situations. Since the model hasn't seen this data during training, this stage helps us to understand how effectively it can estimate house prices for new, unseen cases. Typically, we divide the dataset into training and testing sets randomly, ensuring that both groups share similar characteristics and distributions. A common approach is to assign about 80% of the data for training and 20% for testing. As for the number of training rounds, it is based on the several factors, including the dataset's complexity, the machine learning algorithm selected, enhance the model's precision and optimize its overall performance.

IV. METHODOLOGY

To estimate housing price in this model, we use some machine learning methods- Support vector machine (SVM), random forest, XGBoost, Lasso regression, and linear regression were some of the methods used in our investigation.

Algorithm: During the development of this model, we used Support Vector Machine (SVM), Random Forest, XGBoost, Lasso Regression, and Linear Regression for training. Among these, Random Forest gives the highest accuracy in predicting housing prices. Following closely is the XGBoost algorithm, which we favor for its strength in managing complex, structured datasets and its effective handling of missing values and outliers. As a result, we suggest using XGBoost for real estate price predictions, However, the choice of algorithm may vary based on the specific types and dimensions of the data being used. For our model Random Forest seems to be best choice.

V. IMPLEMENTATION

Here are the steps that we followed in implementation.

5.1 Data Collection

Gather the dataset either from GitHub or it will be also available on Kaggle, that contains important details about houses, like their location, numbers of rooms, square footage, and sale prices. Make sure the dataset includes the features you plan to examine.

5.2 Data Pre-processing

Prepare the collected data for training your model by cleaning and organizing it. Take care of any missing values, apply feature scaling to standardize the ranges, encode categorical variables, and deal with outliers. You might also want to look into feature engineering methods to generate new, valuable features.

5.3 Model Selection

Pick an appropriate machine learning algorithm for forecasting house prices, taking into account the size of dataset, the complexity of the features, and how easy it is to interpret the result. Some popular options include linear regression, decision trees, random forest, and more other methods. We evaluated and tested five different algorithms, and ultimately, XGBoost emerged as the most effective.

5.4 Exploratory Data Analysis

During the exploratory data analysis (EDA) phase of our house price prediction project, we created a visual representation to illustrate how the variables Balcony, bathroom, and price are interconnected.

The insights gained from this EDA can be incredibly useful for prospective homebuyers, real estate professionals, and property developers, as it highlights the key elements that affect house prices.

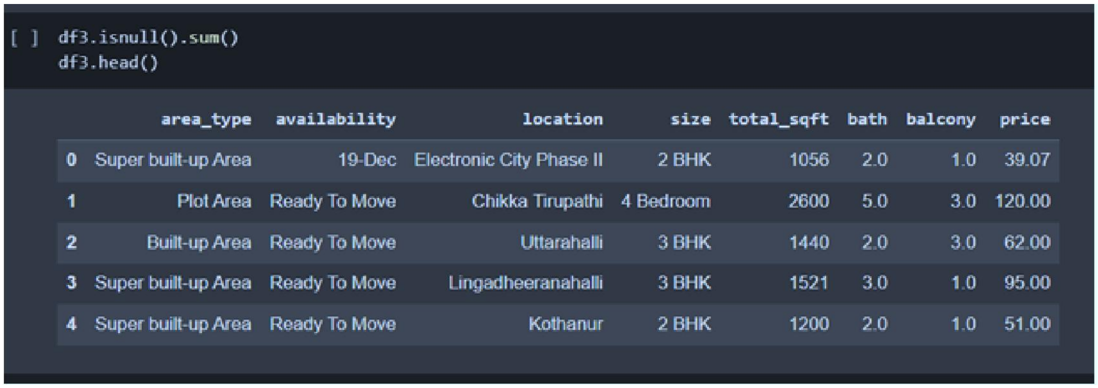
5.5 Correlation Heatmap

As part of our exploratory data analysis (EDA) for predicting house prices, we generated a correlation heatmap to explore the connections between the variables of bathrooms, balconies, and price. This heatmap visually communicates the strength and nature of correlations among these variables. The correlation heatmap provides significant insights into how bathrooms, bedrooms and balconies related to house prices. We found a positive correlation between the number of bathrooms and the home's price, indicating that properties with a greater number of bathrooms tend to be priced higher. Similarly, there was a positive correlation observed between the count of balconies and house prices, suggesting that homes with more balconies might also attract higher selling prices.

5.6 Training and Testing the Model

In our efforts to predict house prices, we took an approach during the training and testing phases of our model by utilizing five distinct algorithms. This method enabled us to assess how well each algorithm performed and its effectiveness in forecasting house prices. The algorithms we used included linear regression, Lesso regression, XGBoost, random forest, and support vector machines (SVM). We trained each algorithm on a segment of the prepared dataset before testing it on another dataset to measure its predictive accuracy.

During the training and testing phases, we fine-tuned the hyperparameters for each algorithm, employing strategies like cross-validation and grid search to enhance their performance. We also used evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared to compare and evaluate the accuracy and predictive capabilities of our trained model.



```
[ ]  df3.isnull().sum()
     df3.head()
```

	area_type	availability	location	size	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	1200	2.0	1.0	51.00

Fig. 2. Data Pre-processing

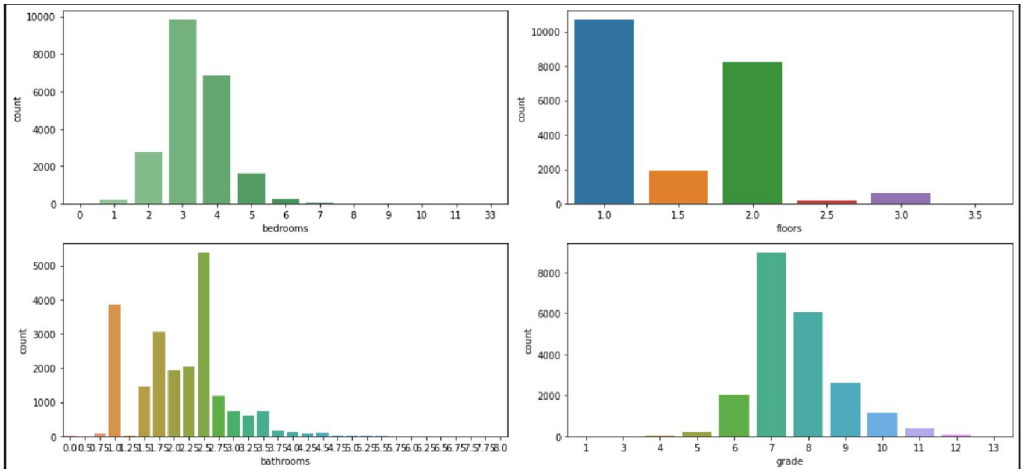


Fig. 3. Exploratory Data Analysis

DOI: 10.48175/IJARSCT-26994

ISSN 2581-9429 IJARSCT

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

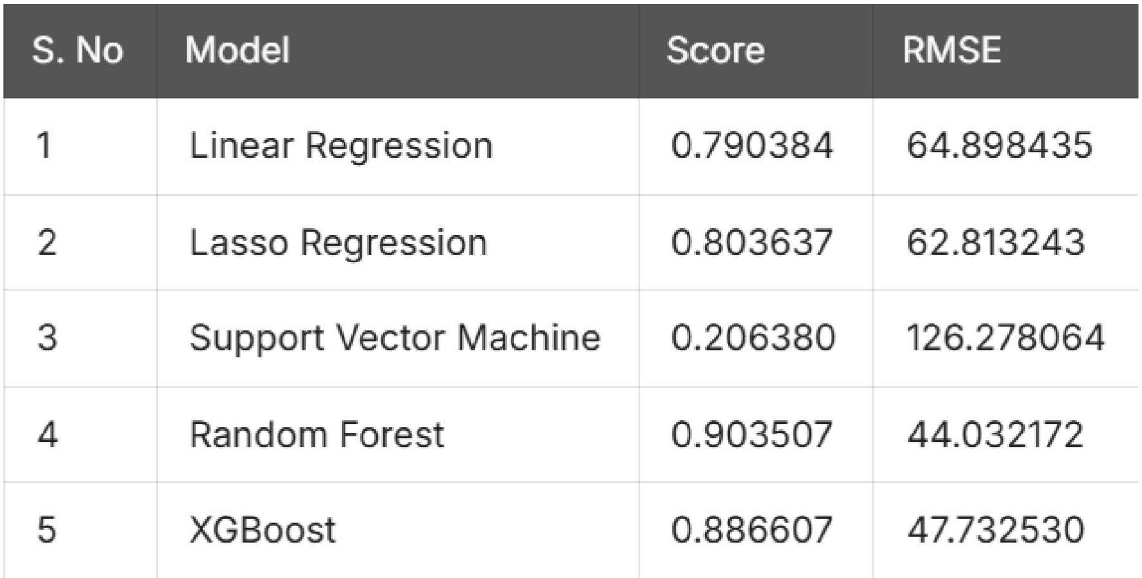
ISSN: 2581-9429

Volume 5, Issue 8, May 2025

Impact Factor: 7.67

VI. RESULTS AND ANALYSIS

To use various machine learning algorithms for solving this problem. Random Forest achieves a high accuracy score of 0.903 and a low root mean squared error (RMSE) value of 44.032. This suggests that the Random Forest model captures the underlying patterns and relationships in the data effectively, resulting in accurate predictions of house prices. Similarly, XGBoost achieves a commendable accuracy score of 0.887 and a reasonably low RMSE value of 47.733. XGBoost is a boosting algorithm that builds an ensemble of weak learners iteratively. It has the ability to handle complex feature interactions and can effectively capture non-linear relationships, resulting in accurate predictions. The regularization techniques employed in XGBoost help prevent overfitting and improve generalization performance.



S. No	Model	Score	RMSE
1	Linear Regression	0.790384	64.898435
2	Lasso Regression	0.803637	62.813243
3	Support Vector Machine	0.206380	126.278064
4	Random Forest	0.903507	44.032172
5	XGBoost	0.886607	47.732530

The outstanding performance of Random Forest and XGBoost can be linked to their proficiency in managing high-dimensional datasets, identifying intricate relationships, and handling feature interactions effectively. These algorithms are recognized for their strength, scalability, and adaptability across various machine learning applications.

VII. CONCLUSION

The main aim of the "House Price Prediction Model Using Machine Learning" project is to estimate house prices based on various features in the data we have. After training and testing the model, we achieved an impressive accuracy of around 90%. To make this model stand out from others, it's important to incorporate additional factors such as taxes and air quality. This way, buyers can make informed decisions within their budgets while reducing the risk of financial loss. We applied several algorithms to assess house values, leading to more precise and accurate selling prices. This is sure to be a great advantage for people looking to buy homes. It's essential to consider and address the many factors that influence housing prices.

REFERENCES

[1]. Available:https://www.researchgate.net/publication/347584803_House_Price_Prediction_using_a_Machine_Learning_Model_A_Survey_of_Literature

[2]. Ankit Mohokar, Nihar Baghat, and Shreyash Mane. House Price Forecasting Using Data Mining, International Journal of Computer Applications. 152:23–26.

[3]. Atharva Chogle, Priyankakhaire, Akshata Gaud, and Jinal Jain. An article titled House Price Forecasting Using Data Mining Techniques was published in the International Journal of Advanced Research in Computer and Communication Engineering. 6:24-28.

DOI: 10.48175/IJARSCT-26994

[4]. Available: https://www.ijraset.com/research paper/house-price-prediction-using-ml

[5]. Available: https://ieeexplore.ieee.org/document/8473231

[6]. Shiva Keertan J, Subhani Shaik. Machine Learning Algorithms Prediction, for International Oil Journal Price of Innovative Technology and Exploring Engineering. 8(8).

[7]. KP Surya Teja, Vigneswar Reddy and Subhani Shaik, Flight Delay Prediction Using Machine Learning Algorithm XGBoost, Jour of Adv Research in Dynamical & Control Systems. 11(5).

[8]. Subhani Shaik, Vijayalakshmi K, Ramakanth Reddy. Location based house of Advanced and prediction using data science techniques", Asian Journal of Advanced Research and Reports. 17(4).

[9]. Jae Kwon Bae, Byeonghwa Park. Housing Price Forecast Using Machine Learning Algorithms. 42:2928–2934.

[10]. Subhani Shaik, Uppu Ravibabu. Classification of EMG Signal Analysis based on Curvelet Transform and Random Forest tree Method. Paper selected for Journal of Theoretical and Applied Information Technology (JATIT). 95.