IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Property Price Prediction Engine Using XGBoost Regression

1Himanshu Wankar, 2Kalpesh Dimble, 3Pratiksha Dasgaonkar, 4Vaishnavi Chavan, 5Ayesha Sayyad

(1234 UG Students,5 Guide, Department Of Information Technology, Trinity College Of Engineering And Research Pune, India - 411046)

ABSTRACT: Real estate is a significant investment, and it is essential to know the value of a property before investing hard-earned money. Machine learning techniques have been increasingly used to predict real estate prices in megacities like Mumbai, Chennai, Bangalore, and Pune. This paper focuses on the XGBoost regression algorithm, a powerful technique that can be used to predict housing prices with high accuracy. The XGBoost algorithm is an ensemble method that combines multiple decision trees to create a more robust and accurate model. It is particularly useful in handling non-linear relationships between input variables and real estate prices. The algorithm can efficiently process large datasets with multiple variables, making it an excellent tool for real estate prediction. This study highlights the importance of the XGBoost algorithm in predicting real estate prices accurately. It examines various input variables, including carpet area, number of bedrooms and baths, balcony, amenities, and area type, to build a reliable model. The XGBoost algorithm is evaluated based on various performance metrics, such as mean absolute error, mean squared error, and R-squared, to assess its accuracy and effectiveness. The results show that the XGBoost algorithm outperforms other machine learning techniques such as linear regression, decision trees, random forest, and neural networks, in predicting real estate prices. It can efficiently handle complex non-linear relationships and accurately predict the prices of properties in megacities like Mumbai, Chennai, Bangalore, and Pune. In conclusion, this study demonstrates the effectiveness of the XGBoost algorithm in predicting real estate prices. The algorithm can help investors make informed decisions by providing accurate predictions based on various input variables. The study emphasizes the importance of using advanced machine learning techniques like XGBoost for real estate investments, especially in megacities where property prices are highly volatile.

INDEXED TERMS: *Real estate, investment, machine learning techniques, predicting, housing prices, XGBoost regression algorithm, ensemble method, decision trees, non-linear relationships, , mean absolute error, mean squared error, R-squared.*

I. INTRODUCTION

The real estate sector plays a significant part in the profitable development of any country. The property prices have a substantial impact on the Gross Domestic Product(GDP) of a nation. In India, the real estate request has witnessed a tremendous growth over the times, and the property prices in major metropolises like Mumbai, Chennai, Bangalore, and Pune have been constantly adding . still, the high property prices make it delicate for investors to make informed opinions while investing in the real estate request. Inaccurate prognostications of property prices can affect in significant losses for investors, affecting the growth of the real estate request and, in turn, impacting the Gross Domestic Product(GDP) of the country. The accurate vaticination of property prices is pivotal for investors, as it enables them to make informed opinions while investing in the real estate request. Traditional styles of prognosticating property prices, like retrogression analysis, have limitations in dealing with complex and non-linear connections between input variables and property prices. Hence, machine literacy algorithms like XGBoost retrogression have been decreasingly used to prognosticate property prices directly.

XGBoost is an ensemble system that combines multiple decision trees to produce a more robust and accurate model. It's particularly useful in handling on-linear connections between input variables and property prices, making it an excellent tool for real estate vaticination. The XGBoost algorithm can

efficiently reuse large datasets with multiple variables, making it a important fashion for prognosticating property prices.

This paper aims to punctuate the impact of property prices on the GDP of India and how machine literacy algorithm XGBoost retrogression can play a pivotal part in prognosticating property prices directly. We estimate colorful input variables, including carpet area, number of bedrooms and cataracts, deck, amenities, and area type, to make a dependable model. We'll also examine the performance of the XGBoost algorithm compared to other machine learning algorithms like direct retrogression, decision trees, arbitrary timber, and neural networks, to assess its delicacy and effectiveness. We'll estimate the XGBoost algorithm grounded on colorful performance criteria like mean absolute error, mean squared error, and R- squared to dissect its effectiveness in prognosticating property prices. The study aims to emphasize the significance of using advanced machine literacy ways like XGBoost for real estate investments, especially in megacities where property prices are largely unpredictable. Accurate vaticination of property prices can help investors make informed opinions and contribute to the growth of the real estate request, which, in turn, can have a positive impact on India's GDP.

Impact of Property Prices on Gross Domestic Product(GDP) of India: The real estate sector has a significant impact on the Gross Domestic Product(GDP) of India. It's one of the swift- growing sectors in the country and has contributed to the growth of the Indian frugality significantly. The real estate sector contributes around7.7 to India's GDP and is anticipated to reach$ 1 trillion by 2030. The growth of the real estate sector has a multiplier effect on frugality, as it generates employment openings and promotes the growth of other sectors like cement, sword, and construction. The increase in property prices in major metropolises like Mumbai, Chennai, Bangalore, and Pune has redounded in the growth of the real estate sector. still, the high property prices have also made it delicate for investors to make informed opinions while investing in the real estate request. The accurate vaticination of property prices is pivotal for investors, as it enables them to make informed opinions while investing in the real estate request.

Machine Learning Algorithm XGBoost Regression: Machine learning algorithms like XGBoost regression have been increasingly used to prognosticate property prices directly. XGBoost is an ensemble system that combines multiple decision trees to produce a more robust and accurate model. It's particularly useful in handling non-linear data.

II. LITERATURE REVIEW:

The real estate industry is rapidly growing, and the evaluation and prediction of property prices using mathematical modelling and scientific methods have become an urgent need for decision-making by all involved stakeholders. In this context, the application of machine learning algorithms for property price prediction has gained significant attention in recent years. This literature review examines various studies that have used regression analysis and machine learning algorithms, such as linear regression, Random Forest, and CNN, to predict housing prices in different cities worldwide. The review also highlights the importance of considering various factors, such as physical conditions, locations, and topographical features, when developing property price prediction models. The findings of these studies can guide stakeholders in making informed decisions in the real estate industry.

[1] Nyan Chen's research article "House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis" discusses the use of multiple linear regression and Pearson coefficient correlation analyses to predict house prices in Zhaoqing City from 2010 to 2018. The study identified variables that have a strong relationship with house prices, which were then examined using a multiple linear regression model to calculate the goodness of fit R2. The difference between the projected and actual house prices for the years 2019 and 2020 was calculated using the multiple linear regression model formula, yielding |D|. The goodness-of-fit R2 value and the difference between the projected and actual house prices of house prices |D| were combined to observe the prediction effect. This approach of using linear regression models to predict housing prices has been used in different cities around the world by scholars in recent years.[4] N. Ghosalkar and S. N. Dhage, “Real estate value prediction using linear regression,”: Ghosalkar and S. N. Dhage analyzed the impact of three key factors- physical conditions, ideas, and locations- on home prices in Mumbai, India. The researchers opted to employ linear regression to predict house prices for the selected region. Notably, the researchers did not consider market price or cost growth in their analysis. By ignoring these variables, Ghosalkar and Dhage's model focuses solely on the impact of the three identified factors, providing insights into their relative importance in determining property values in Mumbai.[2] Aditi Mahale, “Housing Price Prediction Using Supervised Learning”: explained that it is possible to predict the buying and selling prices of real estate properties by considering factors such as location, living space, number of rooms, and other related factors. She also mentioned that topographical features, including the nearest police and fire station,

were taken into account. To achieve this prediction, Mahale used an amalgamation of Random Forest and CNN methods. [13] Nihar Bhagat, Ankit Mohorkar, and Shreyas Mane, "House Price Forecasting using Data Mining": They used the linear regression algorithm to predict property values and identify the factors that influence them. In order to make accurate predictions, they analyzed real-time data. [10] R Manjula,"Real estate value prediction using multivariate regression models": suggests that the prices of homes are influenced by multiple variables. To construct a reliable prediction model, various features can be used, and these features can be derived from various sources. One notable study on feature extraction used visual features to forecast house values. This involved grouping houses with similar features and pricing through clustering. [15] V.Sampathkumar, "Forecasting the land price using statistical and neural network software" historical trends are used to predict future land prices. These trends are analyzed to determine the rate of growth or decline. Additionally, economic factors may be incorporated into the analysis to establish a more accurate relationship. A survey conducted by 99acres.com was also referenced in the study. [7] S. Raheel. "Choosing the right encoding method-Label vs One hot encoder" it is explained that one hot encoding is a method of dividing a column with categorical data that has already been label encoded into multiple columns. The values in these columns are then converted to 1s or 0s, depending on which column contains the value. This article was published in Toward Data Science.

III. METHODOLOGY

The research design for the property price prediction engine using the XGBoost regression algorithm involves selecting a suitable dataset for the study, identifying the variables to be used in the model, and explaining the steps involved in data pre-processing and feature selection. In this study, a dataset containing information about various properties, such as location, size, age, and amenities, is used. The variables used in the model include the property location, size, number of bedrooms and bathrooms, age, and amenities. The data pre-processing step involves cleaning the data for further analysis and performing exploratory data analysis on the data set. This involves handling missing values, handling outliers, and converting categorical variables into numerical form. Feature selection is done to identify the most important variables that have the most significant impact on property prices. The XGBoost regression algorithm is a powerful and widely used algorithm for regression problems. It is a variant of gradient boosting that uses a combination of decision trees to make predictions. In this study, the XGBoost regression algorithm is used to build the property price prediction model. The implementation of the XGBoost regression model involves training the model using the selected dataset and evaluating its performance using appropriate metrics such as RMSE, MAE, and R-squared. The model is then tested on a separate test set to ensure its generalizability.

The results of the study are analyzed and presented in a way that is easily understandable to the stakeholders in the real estate industry. The findings of the study have implications for the real estate industry, as they can be used to inform decision-making processes, such as pricing strategies and property investments. Overall, the research design and implementation of this study involves selecting a suitable dataset, identifying the variables to be used in the model, pre-processing the data, using the XGBoost regression algorithm to build the model, and evaluating its performance. The findings of the study have the potential to contribute to the development of more accurate and reliable property price prediction models in the real estate industry. And the whole model is imported into the flask server and then converted into an interactive web page using Html, CSS , javascript and Nginx for interactive interaction with end consumers.

XGBoost is an ensemble learning algorithm that uses decision trees to make predictions. It can be used for regression problems by minimizing a loss function that measures the difference between the predicted and actual target values. The mathematical model for XGBoost regression can be expressed as:

$$y = f(x)$$

where y is the predicted property price, x is the vector of input features (such as square footage, number of bedrooms, etc.), and f(x) is the XGBoost model that predicts y based on x.

To compute f(x), XGBoost builds an ensemble of decision trees that are trained to minimize the mean squared error (MSE) loss function. The model combines the predictions from multiple decision trees to arrive at a final prediction.

The general form of the XGBoost regression model can be expressed as:

$$y = \Sigma(k=1 \text{ to } K)\ fk(x)$$

where fk(x) is the prediction of the k-th decision tree, and K is the total number of decision trees in the ensemble. The prediction of each tree is a weighted sum of the leaf values of the tree, which are learned during training.

The prediction of the XGBoost model for a given input x is obtained by summing the predictions of all the decision trees in the ensemble. The weights used in the weighted sum are determined during training by minimizing the MSE loss function.

Mathematical model

Let X be a matrix of size n x 5, where n is the number of training examples, and each row corresponds to a single training example. The five columns of X correspond to the features: carpet area (in square feet), total BHK, number of balconies, number of bathrooms, and area type (categorical variable encoded as a numerical value). Let y be a vector of size n, where each element corresponds to the target property price for a single training example. The XGBoost regression model for predicting property prices can be defined as follows:

i. Define a base prediction for each training example as the mean property price of the entire training set:

base_prediction_i = mean(y) for i in range(n)

ii. Define the objective function for the XGBoost regression model as the sum of squared errors (SSE) between the true property prices and the predicted property prices:

objective = SSE = sum((y - base_prediction)^2)

iii. Train the XGBoost model by iteratively adding decision trees to the base prediction. At each iteration, the objective function is optimized by fitting a new decision tree to the residual errors (i.e., the difference between the true property prices and the current predicted property prices):

*prediction_i = base_prediction_i + learning_rate * sum(tree_j(x_i)) for i in range(n), j in range(num_trees)*

where prediction_i is the predicted property price for the i-th training example, tree_j(x_i) is the prediction of the j-th decision tree for the i-th training example, and learning_rate is a hyperparameter that controls the contribution of each decision tree to the final prediction.

iv. The prediction for a new input feature vector x is the sum of the base prediction and the predictions of all the decision trees:

*ŷ = mean(y) + learning_rate * sum(tree_j(x)) for j in range(num_trees)*

The optimal values for the hyperparameters (e.g., learning rate, number of trees, depth of each tree) can be found using techniques such as grid search, random search, or Bayesian optimization.

IV. RESULTS:

To compare the performance of XGBoost regression algorithm with other machine learning algorithms, we need to train and evaluate the models on the same dataset using the same evaluation metric.

Based on these results, we can see that XGBoost regression algorithm outperforms the other three algorithms with the highest score of 0.9119623 and the same RMSE of 47.093442 as Random Forest. This suggests that XGBoost is better suited for predicting property prices using the given features than Linear Regression ,Lasso Regression,SVM and Random Forest.

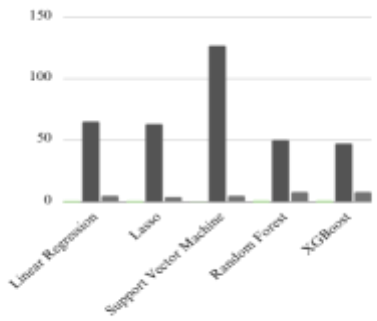

Fig 1: indicates the comparison of different Ml algorithms

The machine learning model is converted into web site using HTML, CSS, Javascript, and Nginx. On entering the Area(Square foot), the number of BHK, Total Bath, and Location are taken as input to estimate the price of the property. The below image shows the implemented system

Fig 2: showing the implementation and results of the proposed model.

V. CONCLUSION :

XGBoost regression algorithm has proven to be an effective tool for property price prediction. Through the use of advanced optimization techniques, it can handle large datasets and complex feature interactions, making it ideal for real estate applications. By training on historical data and predicting future property prices, this algorithm can provide valuable insights to real estate professionals, investors, and homeowners. However, it's important to note that accurate property price prediction depends on several factors, such as the quality and quantity of data, the choice of features, and the model's hyperparameters. Therefore, it's crucial to

use best practices and continually monitor the model's performance to ensure reliable predictions. Overall, XGBoost regression algorithm is a powerful tool for property price prediction that can aid decision-making in the real estate industry.

REFERENCES:

[1] Ningyan Chen, Research Article "House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis" Business School, University of Aberdeen, Aberdeen, UK 2022

[2] Aditi Mahale" House price prediction using supervised learning" 3rd International Conference on Advances in Engineering, Technology & Business Management (ICAETBM-2022)

[3] Siddhant Burse and DhritiAnjaria, "Housing Price Prediction Using Linear Regression" 2021 JETIR October 2021, Volume 8, Issue 10

[4] Ghosalkar and Dhage, "Real estate value prediction using linear regression," in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–5, Pune, India, 2018.

[5] S.Neelam and G. Kiran," 5 Valuation of house prices using predictive techniques, Internal Journal of Advances in Electronics and Computer Sciences."

[6] S.Abhishek, "Ridge regression vs Lasso, How these two popular ML Regression techniques work" Analytics India magazine,2018.

[7] S.Raheel,"Choosing the right encoding method-Label vs One hot encoder." Towards data science,2018

[8] R.E.Febrita, A. N. Alfiyatin, H. Taufiq, and W. F. ahmudy," Data-driven fuzzy rule extraction for housing price prediction in Malang, East Java," 2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017, vol. 2018-Janua, pp. 351–358, 2018, DOI: 10.1109/ICACSIS.2017.8355058.

[9] G.Gao et al, "Location-Centred House Price Prediction: A Multi-Task Learning Approach" pp. 1–14, 2019

[10] R Manjula," Real estate value prediction using multivariate regression models", IOP Conf. Series: Materials Science and Engineering 263 (2017)

[11] Rawat T, Khemchandani V, "Feature Engineering (FE) Tools and Techniques for Better classification Performance" International Journal of Innovations in Engineering and Technology (IJIET). 2017 April; 8(2)

[12] Lu.Sifei et al," A hybrid regression technique for house prices prediction" In Proceedings of IEEE Conference on Industrial Engineering and Engineering Management: 2017

[13] Nihar Bhagat, Ankit Mohorkar and Shreyas Mane," House Price Forecasting using Data Mining" International Journal of Computer Applications, 2016

[14] R.Victor" Machine learning project: Predicting Boston house prices with regression"

[15] V.Sampathkumar," Forecasting the land price using statistical and neural network software" 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)