ISSN: 2347-6540

International Journal of Latest Research In Engineering and Computing (IJLREC)
Volume 11, Issue 1, Page No.64-70, 2024
www.ijlrec.com

HOUSE PRICE PREDICTION: EXPLORING THE IMPACT OF MACROECONOMIC FEATURES BY USING ML

Md. Faiyaz Ansari[1], Anand Raj[2], Khushraj Sinha[3], Saurav Kumar Pathak[4], Mr.Pradeep Chauhan[5]
Student ,CSE, Shivalik College Of Engineering, Dehradun, Uttarakhand, India
mdfaiyaz1303@gmail.com[1]
Student ,CSE, Shivalik College Of Engineering, Dehradun, Uttarakhand, India
reach.anandraj2042002@gmail.com[2]
Student ,CSE, Shivalik College Of Engineering, Dehradun, Uttarakhand, India
khushrajsinha634@gmail.com[3]
Student ,CSE, Shivalik College Of Engineering, Dehradun, Uttarakhand, India
sauravpathak4726@gmail.com[4]
Shivalik College of Engineering ,Pradeep.chauhan@sce.org.in[5]

Abstract :- Accurate prediction of house prices is crucial in the real estate industry, especially with the dynamic nature of the market and the impact of external factors like the COVID-19 pandemic. This research paper aims to explore the effectiveness of machine learning algorithms in predicting house prices, considering both house-specific features and macroeconomic variables to enhance predictive accuracy. The study utilizes datasets containing house-specific features and macroeconomic variables like interest rates, GDP, house price index, CPI, and unemployment rates. Tree-based algorithms such as random forest and XGBoost are employed to develop predictive models. The performance of the models is evaluated using metrics like mean absolute error and mean absolute percentage error to assess predictive accuracy. The results of this research are expected to provide insights into the effectiveness of machine learning algorithms in predicting house prices, highlighting the importance of incorporating macroeconomic variables for enhanced predictive accuracy. By comparing the performance of different algorithms and evaluating the impact of macroeconomic features, this study seeks to contribute to the advancement of house price prediction models, which are crucial for various stakeholders in the real estate market.

Keywords: Gdp, Cpi, Xgboost, Absolute Error

I. INTRODUCTION

House prices are an important reflection of the economy, and housing price ranges are of great interest to both buyers and sellers 1.The general and standardized real estate characteristics are often listed separately from the asking price and general description. Because these characteristics are separately listed in a structured way, they can be easily compared across the whole range of potential houses 1. However, every house also has its unique characteristics, such as a particular view or type of sink, which makes it challenging to estimate an adequate market price 1.Data mining is now commonly used in the real estate market to extract relevant information from raw data and predict house prices, important housing features, and much more 1. Previous research has shown the superiority of tree-based algorithms like random forest and XGBoost in predicting house prices, outperforming traditional methods 123.This research paper aims to explore the effectiveness of machine learning algorithms in predicting house prices, considering both house-specific features and

macroeconomic variables to enhance predictive accuracy 123. The study will utilize datasets containing house-specific features and macroeconomic variables like interest rates, GDP, house price index, CPI, and unemployment rates 13. Tree-based algorithms such as random forest and XGBoost will be employed to develop predictive models, and their performance will be evaluated using metrics like mean absolute error and mean absolute percentage error 123.The results of this research are expected to provide insights into the effectiveness of machine learning algorithms in predicting house prices, highlighting the importance of incorporating macroeconomic variables for enhanced predictive accuracy 123. By comparing the performance of different algorithms and evaluating the impact of macroeconomic features, this study seeks to contribute to the advancement of house price prediction models, which are crucial for various stakeholders in the real estate market 123.

II. LITERATURE SURVEY

The desire for real estate is common among many, seen as a stable investment. However, predicting house prices accurately can be challenging due to various influencing factors. Developing effective prediction models requires thorough research and data analysis, with ongoing efforts by researchers to enhance accuracy and reliability. Advanced techniques like machine learning are being utilized to create robust models for informed decision-making in the dynamic real estate market. "Machine learning based predicting house prices using regression techniques" by J. Manasa, R. Gupta, and N. Narahari, presented at the 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA), focused on utilizing regression techniques for house price prediction 3. "Residential asset pricing prediction using machine learning" by Y. Luo, presented at the 2019 International Conference on Economic Management and Model Engineering (ICEMME), explored the application of machine learning in predicting residential asset prices 3.

"House resale price prediction using classification algorithms" by P. Durganjali and M. V. Pujitha, presented at the 2019 International Conference on Smart Structures and Systems (ICSSS), delved into using classification algorithms for predicting house resale prices 3. "Comprehensive analysis of housing price prediction in Pune using multi-featured random forest approach" by R. Sawant, Y. Jangid, T. Tiwari, S. Jain, and A. Gupta, presented at the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), provided insights into using random forest for housing price prediction 3. "Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism" by P.-Y. Wang, C.-T. Chen, J.-W. Su, T.-Y. Wang, and S.- H. Huang, published in IEEE Access in 2021, introduced a deep learning model for house price prediction incorporating self-attention mechanisms and heterogeneous data analysis 3. These studies collectively showcase the diverse approaches and methodologies employed in the field of house price prediction, ranging from regression techniques to deep learning models, aiming to enhance the accuracy and efficiency of predicting residential property prices.

III . PROPOSED SYSTEM

The proposed system aims to leverage the power of machine learning algorithms, particularly tree-based models, to accurately predict house prices while considering both house-specific features and macroeconomic variables.

XGBoost algorithm:
This is a summary of some of the ways the XGBoost algorithm differs from the gradient boost algorithm. In the original paper about XGBoost, all the steps and calculations have been shown in detail (Chen and Guestrin, 2016). The Extreme Gradient Boosting algorithm, or XGBoost, provides an optimized model known to perform well on various tasks and data sets (Fern´andez-Delgado et al., 2014). It is based on the gradient boosting algorithm, and its greatest advantages include great scalability and speed compared to other gradient boosting techniques.

$$\sum_{i=1}^{n} \mathcal{L}(y_i, F(x_i)) + \gamma T + \frac{1}{2}\lambda O^2. \quad (1)$$

Finally, the prediction of the decision tree is calculated as

$$F_m(x) = F_{m-1}(x) + \eta \sum_{j=1}^{J_m} O_{J,m} I(x \in R_{j,m}) \quad (2)$$

which is the same as in the original gradient boosting algorithm.

RANDOM FOREST

Random forest is a machine learning method that falls into the category of parallel ensembling methods, which will be explained now. It builds upon a concept called bagging, a way to train parallel trees. In bagging, the trees are built with data generated by bootstrap aggregating, which draws multiple random samples with replacements from the original data set. Each subsample B is used to train each tree ˆf. Hence, each tree learns from multiple, slightly different subsamples. (Breiman, 1996)

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x). \quad (3)$$

MEAN ABSOLUTE ERROR (MAE) :

Mean absolute error (MAE) is a measure used to evaluate the performance of regression models. Specifically, it is the mean absolute difference between a given data set's actual and predicted values. It is a popular measure as it is easier to interpret than similar measurements like RMSE (Willmott and Matsuura, 2005).

Let the observed value be denoted by yi and the predicted value by pi . The residual of the model is then calculated as

$$\epsilon = yi - pi . \quad (4)$$

Then, the MAE is calculated as

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\epsilon_i|. \quad (5)$$

Mean absolute percentage error (MAPE)

Mean absolute percentage error (MAPE) is a measure used to calculate the performance of a regression model. The measurement aims to evaluate the model's performance by calculating the relative difference between the predicted and actual values. Therefore, MAPE is suitable when comparing models where the dependent variable has different units. It is also useful when briefing someone unfamiliar with the units being used on a model's performance, and it gives a universal idea of how good a model is. (1) shows the calculation of MAPE, where ϵi is the residuals shown in (13)(Swamidass, 2000).

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\epsilon_i}{y_i} \right| \times 100 \quad (6)$$

K-FOLD CROSS VALIDATION

When training and evaluating a model, it is common to use an error metric such as MAE that estimates how well the model performs on unseen test data. However, when minimizing the in-sample error, the model risks overfitting the to training data set. To avoid this, k-fold cross-validation can be used. K-fold cross-validation estimates the out-of-sample MAE by splitting the training data set into k different groups, using one of them as a validation set, similar to a test data set. The model is trained on the first k − 1 parts of the data and then evaluated on the validation set to give the validation

MAE. This is repeated until all k groups have been used as validation sets. The k-fold cross-validation MAE is then calculated as

$$MAE_k = \frac{1}{k} \sum_{i=1}^{k} MAE_i.$$

7

Data Science- Data wisdom is the first stage in which we take the dataset and will do the data drawing on it. We'll do the data drawing to make sure that it provides dependable prognostications. Machine Learning- The gutted data is fed into the machine literacy model, and we do some of the algorithms like direct retrogression, retrogression trees to test out our model.

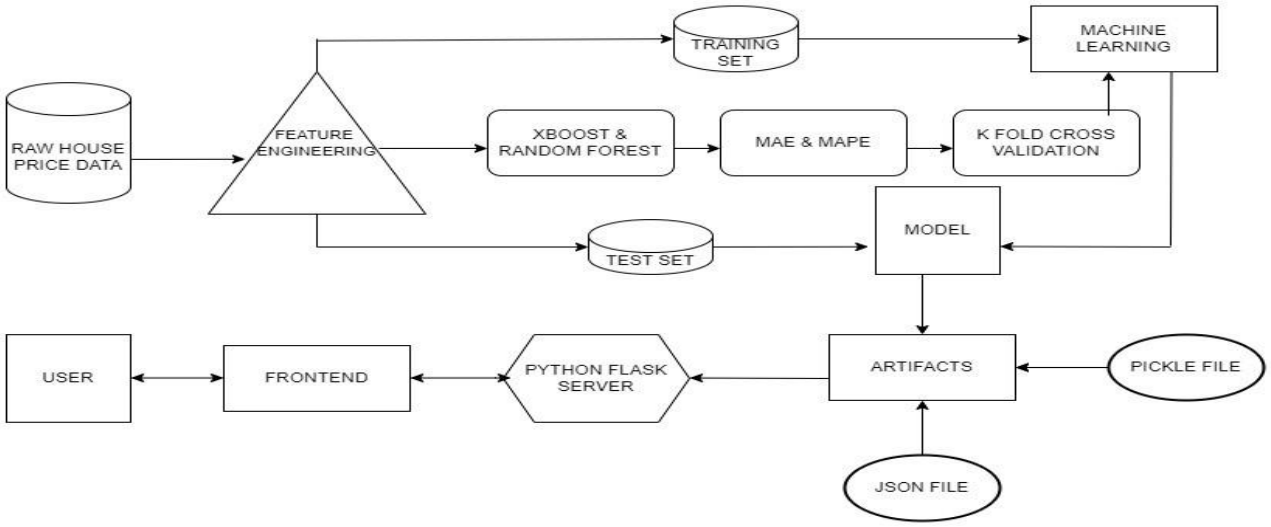

Fig 1.1 ARCHITECTURE

The gutted data is fed into the machine literacy model, and we do some of the algorithms like direct retrogression, retrogression trees to test out our model. Front End (UI) - The frontal end is principally the structure or a figure up for a website. In this to admit an information for prognosticating the price. It takes the form data entered by the stoner and executes the function which employs the prediction model to calculate the predicted price for the house.

IV. DATA VISUALIZATION

Visualization gradually makes complex data more accessible, reasonable, and usable as shown in Fig 2 and Fig 3. Dealing with, analyzing, and transmitting this data presents good and orderly challenges for data representation. This test is addressed by the field of data science and experts known as data scientists. In Fig 2 below shows the scatterplot of price_per_sqft vs Total Square feet of the random place from the dataset Hebbal where blue dot represents 2BHK and green plus represents 3BHK. This plot is with the outliers present in the dataset.

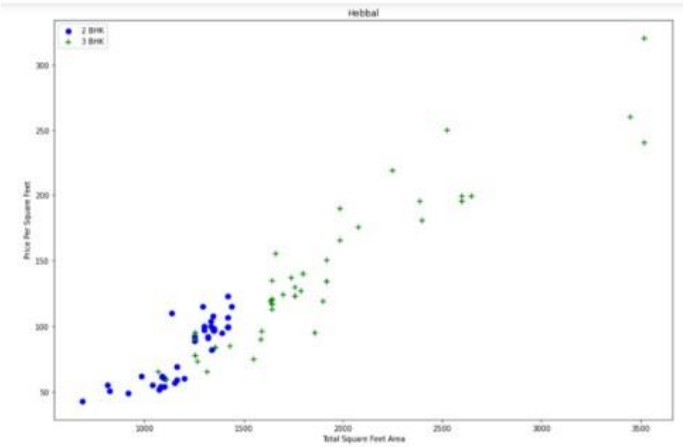

Fig 2: Price Outliers for a place (Hebbal)

In Fig 3 below shows the scatterplot of price_per_sqft vs Total Square feet of a random place from the dataset Hebbal where blue dot represents 2BHK and green plus represents 3BHK. This plot is after removing the outliers present in the dataset by using the function. Also in the above fig we can find one or two green plus which is 3BHK and still shows as outlier after the function is applied. But that is a minor difference where is has come due to the place and its area where the house is present.

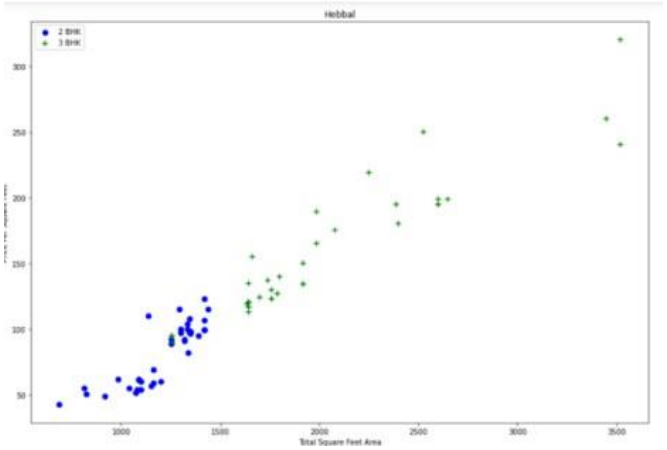

Fig 2: Price after Outliers removed (Hebbal)

A correlation matrix is just a simple visual representation table that gives correlation between the different variables of the table. The matrix gives almost all the possible correlation between the variables possible. Whenever the large datasets are considered it is best option to display the summary of the different patterns of the data. The correlation matrix has the value ranging between -1 to +1. Thus the positive number shows the positive links among the variables while the negative number shows the negative link between the variables that are considered. In the Fig 4 below five variables (features-total_sqft, bath, price, bhk, and price_per_sqft) are plotted and the correlation among them is displayed. For Heatmap the Python library sns is used for data visualization that is based on matplotlib.

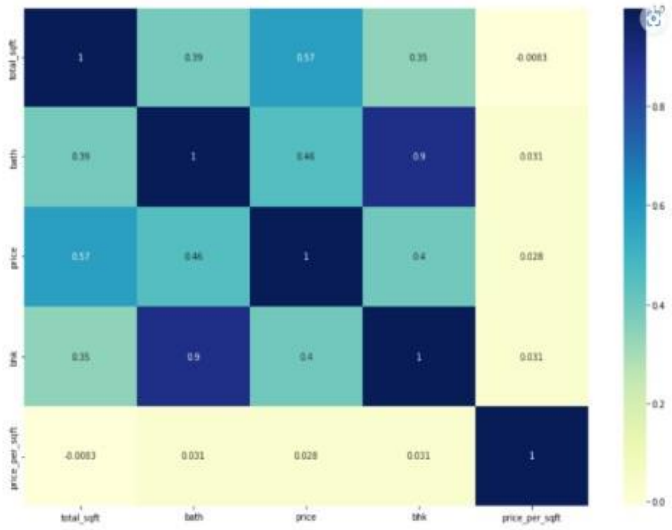

Fig 4 Correlation Matrix

V: RESULTS

The above Fig 5 shows the comparison between the various algorithms used to build the price prediction model, where it is found out that the XBOOST algorithm gives the maximum accuracy of about 84.77 percent. While other algorithms Random Forest and Decision Tree gives 72.26 and 73.16 percent respectively.

	model	best_score	best_params
0	XBOOST	0.847796	{'normalize': False}
1	Random Forest	0.726745	{'alpha': 2, 'selection': 'random'}
2	Decision Tree	0.731685	{'criterion': 'mse', 'splitter': 'random'}

Fig 5 Comparison of the accuracy

V. CONCLUSION

In this study, various machine learning algorithms are used to estimate house prices. All of the methods were described in detail, and then the dataset is taken as input, applied the various models to give out the results of the prediction. The presentation of each model was then compared based on features where it is found that Xboost algorithm gives maximum accuracy of about 84 to 85% after a proper comparison with decision tree and Random Forest. The correlation matrix also displays the visualization of the larger data into compact pattern. Thus the model can work with decent efficiency giving the required features to the customer.

REFERENCES

[1] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.

[2] N. Apergis et al. Housing prices and macroeconomic factors: prospects within the european monetary union. International Real Estate Review, 6(1):63–74, 2003..

[3] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia, 2018, pp. 35-42, doi: 17.1109/iCMLDE.2018.00017.

[4] M. Jain, H. Rajput, N. Garg and P. Chawla, "Prediction of House Pricing using Machine Learning with Python," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 570-574, doi: 10.1109/ICESC48915.2020.9155839.
[5] Nihar Bhagat, Ankit Mohokar and Shreyash Mane. House Price Forecasting using Data Mining. International Journal of Computer Applications 152(2):23-26, October 2016..
[6] J. Manasa, R. Gupta and N. S. Narahari, "Machine Learning based Predicting House Prices using Regression Techniques," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 624-630, doi: 10.1109/ICIMIA48430.2020.9074952.
[7] J. Ekberg and L. Johansson. Comparison of different machine learning methods' capability to predict housing prices. Student essay, Royal Institute of Technology, Diva, 2022.
[8] J. Gareth, D. Witten, T. Hastie, and R. Tibshirani. An introduction to statistical learning: with applications in R. Spinger, 2013.
[9] N. N. Ghosalkar and S. N. Dhage, "Real Estate Value Prediction Using Linear Regression," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697639.
[10] T. Xu. The relationship between interest rates, income, gdp growth and house prices. Research in Economics and Management, 2(1):30–37, 2017.