MACHINE LEARNING APPROACH FOR HOUSE PRICE PREDICTION

Noone Srikanth
Master of Computer Applications, Chaitanya Bharathi Institute of Technology(A), Hyderabad, Telangana, India
noonesrikath@gmail.com

Dr. M. Ramchander,
Assistant Professor, Master of Computer Applications, Chaitanya Bharathi Institute of Technology(A), Hyderabad, Telangana, India

ABSTRACT

To date, there's limited research on using machine learning to predict property values in India, despite the real estate market's constant price fluctuations. Machine learning can help us understand and forecast these changes more accurately. This, in turn, can assist both buyers and sellers in making informed decisions in the real estate market. The primary goal of this project is to predict house prices by considering various real-world factors. We aim to evaluate different parameters that influence prices. To simplify large datasets and identify the most critical factors for predicting house prices, we'll use various methods to select the most relevant features. This will enable us to make more precise predictions about property values.

KEYWORDS: Regression, Machine Learning, Feature Selection, Price Prediction.

I.INTRODUCTION

The Indian real estate market is renowned for its intricacies and regional dynamics, offering a diverse landscape for property transactions. This study focuses specifically on the house selling market in Hyderabad, a prominent city in India. Hyderabad boasts a thriving real estate sector, encompassing residential, commercial, and industrial properties. Understanding the nuances of the house selling market in Hyderabad is essential for the development of an accurate prediction model tailored to the local population's needs.

The primary objective of our project is to construct a machine learning model capable of providing precise property price predictions to the general public, thereby bridging the information gap between buyers and sellers. The presence of intermediaries in real estate transactions often results in inflated prices, posing challenges for buyers seeking equitable deals. By delivering accurate property price predictions, our model aims to eliminate the necessity for mediators and establish a more transparent and efficient market for both buyers and sellers in Hyderabad.

While previous studies in the literature have typically focused on a limited set of features, our research adopts a comprehensive approach, considering 24 distinct features. These features encompass a broad spectrum of factors influencing property prices, including location, size, amenities, infrastructure, market trends, and socio-economic indicators. By incorporating this diverse set of features, our model strives to capture the intricate nature of the Hyderabad house selling market and enhance the accuracy of price predictions.

To achieve our research objectives, we employ various regression techniques, including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. These techniques have demonstrated their effectiveness in predicting continuous variables, such as property prices, by analyzing the relationships between input features and target variables. By deploying multiple regression models, we can compare their performance and select the most suitable approach for accurately predicting house prices in Hyderabad.

In assessing the prediction models' performance, we consider several metrics, including R2 (coefficient of determination), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error). These metrics provide valuable insights into the precision and correctness of our models when estimating property values. R2 evaluates the predictability of the target variable's variance from the input features, while MAE and MAPE gauge the average error magnitude in predictions, accounting for both absolute and relative disparities.

By recognizing the distinctive characteristics of the house selling market in Hyderabad and incorporating a wide array of features, our research aims to contribute to the development of a reliable and accurate prediction model for property prices. The successful implementation of such a model can bridge the information gap between buyers and sellers, empowering ordinary individuals to make informed decisions and obviating the need for intermediaries. Furthermore, the utilization of feature selection methods, multiple regression techniques, and comprehensive evaluation metrics enhances the robustness of our models and ensures the provision of reliable price predictions in Hyderabad's real estate market.

II.LITERATURE SURVEY

The literature review offers a comprehensive overview of prior research in the field of property price prediction, showcasing various methodologies and findings from distinct studies.

An investigation conducted by Saiyam Anand et al. [1] successfully predicted house prices by considering four independent factors: location, square footage, number of bedrooms (bhk), and number of bathrooms. Their study highlighted the dependency of property prices on these factors, resulting in accurate predictions and a functional model, particularly in the context of Bengaluru.

In the work of Peng et al. [2], a potent algorithm known as XGboost was employed to forecast the prices of secondhand houses in Chengdu, China. The study revealed XGboost's superior performance in handling intricate data patterns, avoiding overly simplistic predictions often associated with decision trees or linear regression models.

Madhuri et al. [3] conducted research employing various regression techniques to predict house prices. These techniques proved beneficial in assisting sellers in determining optimal selling prices for their properties while furnishing buyers with precise pricing information.

Mu et al. [4] conducted a comparative analysis of two distinct methods, Support Vector Machine (SVM) and Least Squares SVM, for house price prediction. Their findings indicated the superior performance of both methods over the commonly used Partial Least Squares technique.

Poursaeed et al. [5] introduced the idea that a property's interior and exterior appearance significantly influences its price. To explore this concept, they developed a unique model incorporating images of various house features and conducted experiments using real estate databases such as Zillow, Redfin, and Trulia.

M. Ceh et al. [6] employed a specialized machine learning technique, the random forest, to predict real estate sales. They compared the outcomes of this method with those of the traditional HPM (House Price Model). The study concluded that the random forest approach outperformed the traditional method, demonstrating its superiority in sales prediction.

In a study conducted by T. Dimopoulos et al. [7], two methods, Random Forest (RF) and Linear Model Regression (LMR), were compared for predicting apartment prices in the Nicosia area of Cyprus, utilizing real estate data. The results highlighted the random forest approach's enhanced accuracy in price prediction.

This literature review provides valuable insights into previous research efforts, offering a foundation for our study's approach and methodology. The following sections will elucidate our unique contributions and the methodology applied to predict house prices in Hyderabad.

III.METHODOLOGY

A. DATASET

The dataset [8] used in this project was obtained from Kaggle, a well-known website. It contains a vast amount of information with over 2434 rows and includes 24 attributes in which each and every feature can impact the outcome. It consists of various features like No. of bed rooms, Area in sqft, Resale or new house and other facilities.

B. DATA CLEANING AND PREPROCESSING

Getting the data ready for analysis and modeling is a crucial step, but it can be challenging. The unprocessed data cannot be used directly because ML algorithms need numbers not words or missing information. Different algorithms have specific requirements for the data they can work with. So, as a part of the data cleaning process unnecessary columns and any unidentified values are dropped or removed. As part of the preprocessing steps, the location category is converted to numbers using mean encoder, and the data scaling is done to make sure everything is on a similar scale. All other features which are having yes/no values are also converted to 0's and 1's in-order train the model. These adjustments help to find meaningful patterns and relationships in the data.

C. VISUALIZATION

Data visualization turns numbers and information into easy-to- understand pictures and graphs. Being able to quickly identify patterns and trends in data not only makes it more engaging but
also facilitates better decision-making

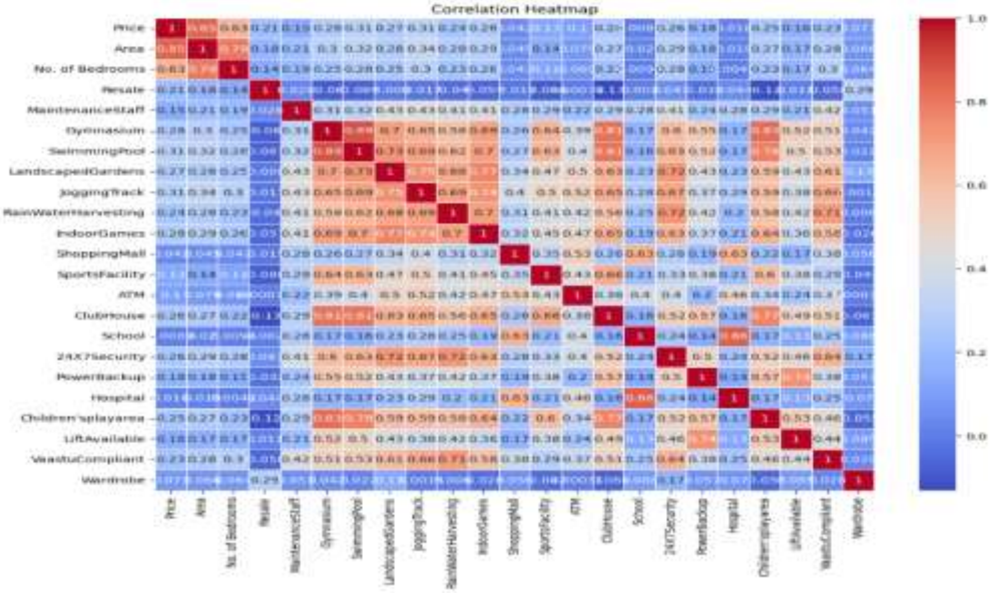

Figure 1. heatmap with correlation scores

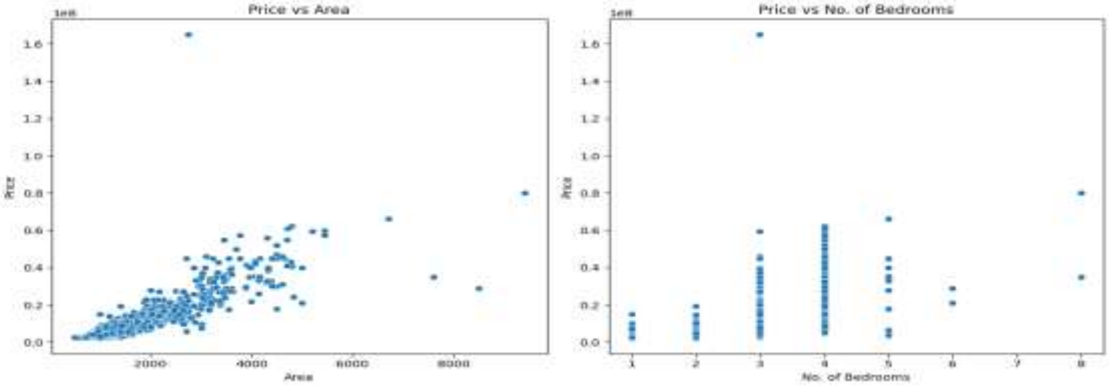

Figure 2. Correlation between Area & No. of bed rooms

Here in my project, I utilized heat map visualization to explore the relationships between 24 different features. By employing a color-coded representation, the heat map effectively showcased the strength and direction of correlations among the features. Also, I employed scatter plot visualization to examine the relationship between the variables and highlight the highly correlated features with the price variable. By plotting the Area and Number of bedrooms (which are highly co-related) against the price, I could visually depict the patterns and trends in the data. The scatter plot showcased how changes in the Area and Number of bedrooms affected the price variable.

D. FEATURE SELECTION

Feature selection is like picking out the most important puzzle pieces that help solve a problem in machine learning. It's about finding the key factors that have the biggest impact on predicting outcomes accurately. By selecting the right features, we can simplify the model and focus on what really matters, making our analysis more efficient and effective. Wrapper methods are a popular approach to feature selection that involve evaluating subsets of features using a specific machine learning algorithm. One common wrapper method is the sequential feature selection. Sequential feature selection is a systematic procedure that iteratively adds or removes features based on their impact on model performance. It starts with an empty set of features and gradually selects or eliminates features until a stopping criterion is met. The selection or elimination is determined by the performance of the model on a validation set or through cross-validation.

IV MACHINE LEARNING MODELS

In this study, I employed various machine learning algorithms to predict house prices. The algorithms used in this analysis include: Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor and XGBoost Regressor. To assess the performance of these models, I utilized the sci-kit learn Python library. Then evaluated the models using several performances

metrics, which includes R-square, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE). These metrics provide valuable

insights into the accuracy and effectiveness of the models in predicting

house price.

A. Linear Regression

LR is a supervised ML technique used for regression tasks, where it operates under the assumption of a linear connection between an input variable (x) and a solitary output variable (y). By incorporating multiple independent features from our dataset. Multiple Linear Regression (MLR) enabling us to estimate the correlation between two or more independent variables and a dependent variable, considering the potential dependence of prices on these diverse features.

B. Random Forest Regressor

RF is like a team of models working together to make more accurate predictions. Instead of relying on just one model, it combines multiple models to create a stronger and more reliable model. Here is how it works: Each model in the random forest is like a decision tree, where it makes decisions based on different factors. However, what makes random forest unique is that each tree uses a different subset of features from the dataset. This helps to create a diverse set of decision trees that are not strongly correlated with each other. By combining the predictions of these individual decision trees, the random forest algorithm produces a final predicted result. This ensemble of models reduces the chances of overfitting or relying too much on a single model's bias.

C. Decision Tree Forest Regressor

DT is a powerful algorithm used in machine learning for making predictions. It operates by constructing a tree-like model of decisions and their possible consequences. Each decision tree in the ensemble learns from a different subset of features from the dataset, allowing for a diverse set of decision trees to be created. The algorithm makes predictions by aggregating the predictions of individual decision trees. This ensemble approach helps reduce the risk of overfitting and minimizes the impact of any particular decision tree's biases. The Decision Tree Regressor provides an effective and reliable method for making accurate predictions in various domains of machine learning.

D. Gradient Boosting Regressor

GB Regressor is an ML algorithm that is used for making predictions, especially when we want to predict numerical values (regression tasks). It is a powerful and popular algorithm renowned for its capability in handling intricate patterns within the data. Gradient Boosting Regressor is an algorithm that builds a series of models, each one correcting the errors of the previous models to make accurate predictions.

E. XGB Regressor

Extreme Gradient Boosting Regressor is an ML that creates a powerful predictive model by combining many weak models together. It works by repeatedly improving the weak models' performance based on their errors, allowing them to learn from each other and make better predictions collectively.

V.RESULT ANALYSIS

After assessing the results of various ML algorithms on the dataset using different algorithms, we compared various metrics. These metrics include R-Squared Score (R2_score), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE). By taking these metrics into consideration, it becomes possible to assess and compare the performance of various machine learning models.

The performance metrics are defined as follows:

R-Squared Score (R2_score):

The R2 score assesses the alignment between the model's predictions and the real values of the target variable. This metric quantifies the extent to which the model explains the variance within the dataset. The R2 score has a scale of 0 to 1: 0 signifies the model's inability to grasp any of the fluctuations in the target variable, while 1 signifies the model's precise prediction of the target variable with no inconsistencies.

R2 score = 1 - (SS_res / SS_tot) --------------------------------------- (1)

SS_res stands for the sum of squared residuals, which assesses the squared variance between predicted and observed results. Conversely, SS_tot signifies the total sum of squares, measuring the squared variability between actual data points and the mean value.

Mean Absolute Error (MAE):

By performing the process of calculating the mean, we employ MAE as a metric for assessing the average absolute variance between predicted and observed values within a regression scenario. The derivation of MAE involves computing the mean value of these absolute variances, resulting in a comprehensive quantification of the model's predictive deviations from the true values. A decreased MAE value corresponds to heightened precision and fidelity of the model's predictions.

$\text{MAE} = (1/n) * \Sigma|i=1 \text{ to } n| (|y_i - \hat{y}_i|)$ ------------------------------------ (2)

Mean Absolute Percentage Error (MAPE):

MAPE is a widely adopted metric for assessing prediction model accuracy. It calculates the mean percentage difference between predicted and actual values. This measure offers a relative evaluation of prediction error, proving especially valuable when handling datasets with diverse scales and magnitudes.

$\text{MAPE} = (1/n) * \Sigma(i=1 \text{ to } n) |(y_i - \hat{y}_i) / y_i| * 100$ ----------------------- (3)

Table 1. Performance metrics of Regression Models with all features

Model	R2 Score	MAE	MAPE
XGB	0.9343233	1022401	10.342847
Gradient Boosting	0.9306369	1228389.4	12.948701
Random Forest	0.9153654	1094763.3	10.330984
Decision Tree	0.8751546	1214068.1	12.234033
Linear Regressor	0.8389191	1771848.7	19.055801

Table 2. Performance metrics of Regression Models with 15 features

Model	R2 Score	MAE	MAPE
XGB	0.93297628	1149839.3	0.111847
Gradient Boosting	0.91324414	1232632.1	0.128569
Random Forest	0.91521249	1127731.1	0.104290
Decision Tree	0.88537491	1191815.9	0.116166
Linear Regressor	0.84176613	1764759.9	0.187265

Table 3. Performance metrics of Regression Models with 10 features

Model	R2 Score	MAE	MAPE
XGB	0.93192217	1108839.4	0.110227
Random Forest	0.90974611	1136293.7	0.104473
Gradient Boosting	0.92061409	1216389.1	0.121840
Decision Tree	0.88604653	1223317.6	0.118367
Linear Regressor	0.84203251	1766546.9	0.187708

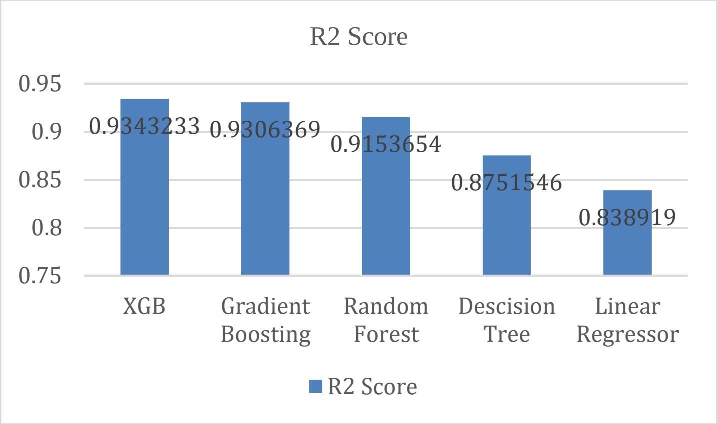

Figure 3. Comparison of R2 score with all features

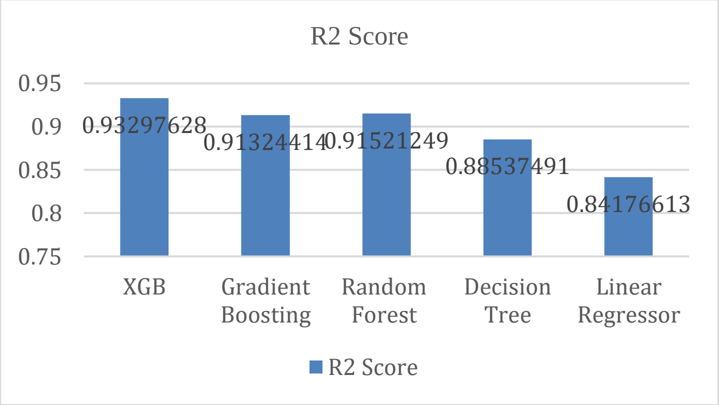

Figure 4. Comparison of R2 score with top 15 features

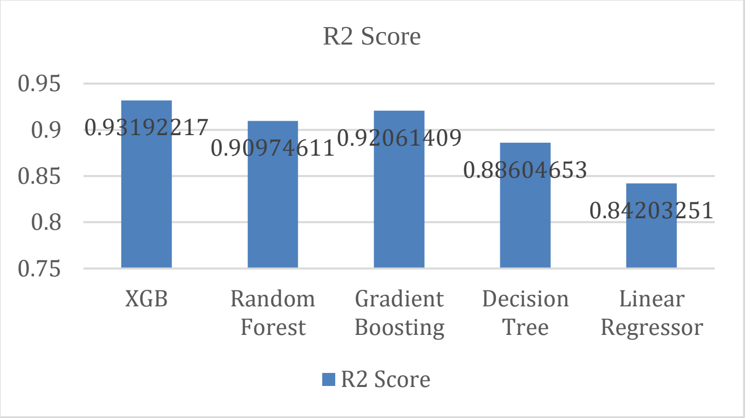

Figure 5. Comparison of R2 score with top 10 features

VI. CONCLUSION AND FUTURE SCOPE

In this paper, I aimed to predict house property prices using various machine learning algorithms and compared them in terms of performance metrics. The machine learning algorithms includes Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression and XGB Regression. All these algorithms were trained on a dataset containing 2434 records with 24 attributes.

After evaluating the performance metrics of all these algorithms, it was observed that the XGBoost Regressor performed exceptionally well in terms of performance metrics, achieving the highest adjusted R-squared value of 0.9343233, the lowest MAE of 1022401 and MAPE of 10.342847 surpassing the other models. This indicates that the XGBoost Regressor algorithm is exceptionally effective in predicting house prices based on the given dataset with all features without applying feature selection methods. Also I have applied sequential feature selection method to select most important features in predicting the house prices by adjusting number of features 15, 10 and 5 to compare. This is also done by applying all the above-mentioned machine learning models by changing the number of features count to observe the performance difference. Overall, again the XGBoost regressor has performed well followed by Random Forest regressor with good results.

For enhancement there is need of adding some more features which will change the house price prediction results. The features like - on which floor the house is present, railway station and other transportation availability etc., can be added. By adding these features will significantly changes the prediction. And it shows good results in prediction as these facilities impacts the house prices undoubtedly.

VII. REFERENCES

[1] Saiyam Anand, “Real Estate Price Prediction Model”, 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) | 978-1-6654-3811-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAC3N53548.2021.9725772
[2] Z. Peng, Q. Huang, and Y. Han, ‘‘Model research on forecast of secondhand house price in Chengdu based on XGboost algorithm,’’ in Proc. IEEE 11th Int. Conf. Adv. Infocomm Technol. (ICAIT), Oct. 2019, pp. 168–172.
[3] C. R. Madhuri, G. Anuradha, and M. V. Pujitha, ‘‘House price prediction using regression techniques: A comparative study,’’ in Proc. Int. Conf. Smart Struct. Syst. (ICSSS), Mar. 2019, pp. 1–5
[4] J. Mu, F. Wu, and A. Zhang, ‘‘Housing value forecasting based on machine learning methods,’’ Abstract Appl. Anal., vol. 2014, pp. 1–7, Aug. 2014.
[5] O. Poursaeed, T. Matera, and S. Belongie, ‘‘Vision-based real estate price estimation,’’ Mach. Vis. Appl., vol. 29, no. 4, pp. 667–676, May 2018.
[6] M. Ceh, M. Kilibarda, A. Lisec, and B. Bajat, ‘‘Estimating the performance of random forest versus multiple regression for predicting prices of the apartments,’’ ISPRS Int. J. Geo-Inf., vol. 7, p. 168, Oct. 2018.
[7] T. Dimopoulos, H. Tyralis, N. P. Bakas, and D. Hadjimitsis, ‘‘Accuracy measurement of random forests and linear regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus,’’ Adv. Geosci., vol.

45, pp. 377–382, Nov. 2018.

[8] RUCHI BHATIA, “Housing Prices in Metropolitan Areas ofI India”,https://www.kaggle.com/datasets/ruchi798/housing-prices-in-metropolitan

[9] Dr. M. Ramchander and Dr. Lakshi Sreenivasareddy.D , “A Model for Improving Classifier Accuracy using Outlier Analysis Methods”, Artificial Intelligence and Machine Learning (AIML) Journal, ISSN:1687-4846, Delaware, USA, December 2015.

[10] M. Ramchander, Dr. Y. Rama Devi, Dr. Lakshi Sreenivasareddy.D ,“Cluster Sampling to Improve Classifier Accuracy in Continuous data” The international journal of analytical and experimental model analysis Volume XIII, Issue VI, June/2021 ISSN NO:0886-9367.