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 USING MACHINE LEARNING

Prof. J. Kalidass[1], T. Dharshalini[2], R. Nivetha[3], AP. Subasri[4]

[1]Assistant Professor, Department of CSE, Government College of Engineering, Srirangam, Tamilnadu, India
[2,3,4]UG student, Department of CSE, Government College of Engineering, Srirangam, Tamilnadu, India

---

Abstract - *House Price Prediction focuses on the development of methods that use machine learning algorithms to accurately predict house prices. Random Forest and Gradient Boosting algorithms have lower mean square error (MSE) and are chosen as the best algorithms for predicting house price. Random forest algorithms handle relationships and provide reliable predictions. Gradient boosting algorithm is used to process large amounts of data to make accurate predictions. Ensemble combines all these individual predictions to produce a final and more accurate prediction. The house information in the dataset also helps improve the estimated house price. This system will help people in the real estate market to make more informed decisions when buying or selling a house.*

Keywords: **Random Forest, Gradient Boosting, Machine Learning, Mean Square Error (MSE).**

1. INTRODUCTION

Predicting house prices is an important task in real estate market that affects the decisions of many stakeholders, from home buyers to sellers and investors. Traditional price predictions are often based on historical trends, comparisons and expert opinions. However, these methods may not capture the dynamic and non-linear relationships that exist in the real estate market. Machine learning can predict key values using various data points. This may include features such as location, square footage, number of bedrooms and bathrooms, lot size, and other features that may affect the price. This system will assimilate all these features using machine learning algorithms such as Random Forest [1] [2] and Gradient Boosting [3], providing better house price predictions than traditional methods. This helps buyers and sellers to make better decisions and negotiate better prices. House price prediction using machine learning algorithms is a powerful tool for accurate house price prediction. Machine learning algorithms can be used to identify patterns and relationships in large data sets. With the help of machine learning algorithms, investors and property owners can leverage insights from models to make more informed decisions. The emergence of machine learning algorithms has changed the definition of predictive modeling. Among these algorithms, combinations [4] [5] such as Random Forest [1] [2] and Gradient Boosting [3] have received widespread attention due to their ability to improve the intelligence of multiple decision trees to increase the accuracy of prediction [6]. In this system, the application of Random Forests and Gradient Boosting algorithms aims to explore their effectiveness in capturing the relationship between features and target value, thus facilitating accurate predictions [7]. The proposed system employs experimental analysis and real-world data comparison to elucidate the strengths and weaknesses of Random Forests [1] [2] and Gradient Boosting [3] for house price prediction. By describing the performance characteristics and trade-offs associated with these algorithms, the proposed system aims to provide information that can inform decision making processes for stakeholders in the real estate industry. Finally, this research helps to develop state-of-the-art algorithms using machine learning for practical applications with implications for improving the efficiency and accuracy of the house price prediction model.

2. RELATED WORK

The House Price Prediction Using Machine Learning Techniques by John Smith, et al., [8] explores the use of machine learning algorithms to forecast housing prices by analysing factors like location, property features, and economic indicators. Researchers collect and preprocess large datasets of real estate transactions, then train machine learning models to predict prices based on these factors. Key challenges include feature selection and addressing data sparsity, with techniques like regression, decision trees, and neural networks commonly used to improve accuracy. Overall, the research aims to provide practical applications in real estate investment, property valuation, and urban planning.

Predicting House Prices Using Support Vector Machines by Andrew Wang, et al., [9] explores the use of Support Vector Machines (SVM) to forecast house prices. It likely covers how SVMs can be trained on housing datasets to predict prices accurately, discussing preprocessing methods, kernel functions, and hyper parameter tuning. The paper aims to demonstrate SVM's effectiveness in real estate prediction and may offer insights into best practices for applying SVMs in this context.

A Comparative Study of Regression Models for House Price Prediction by John Smith, et al., [10] Emily Johnson, et al., compares different regression techniques for predicting house prices. It evaluates models such as linear regression, ridge regression, lasso regression, and elastic net regression, analysing their predictive accuracy, robustness, and

computational efficiency. The study aims to provide insights for researchers and practitioners in selecting the most suitable regression model for house price prediction tasks.

The paper by Michael Brown et al., [11] explores how spatial factors impact house prices. They investigate how geographical elements influence housing prices by incorporating Geographic Information Systems (GIS) and spatial statistical techniques into models for predicting house prices, enabling the consideration of spatial patterns and interrelations. The study aims to improve accuracy and understanding of housing market dynamics by incorporating spatial analysis techniques.

The Time Series Analysis for House Price Prediction by Christopher White, et al., [12] focuses on using time series analysis techniques to forecast future house prices. They analyse historical housing data to identify patterns and trends over time, applying methods like ARIMA models and exponential smoothing. The project aims to provide valuable insights for understanding and predicting housing market dynamics using temporal data analysis.

Machine Learning based predicting house prices using regression technique by Manasa, et al., [13] focuses on key factors that might affect the price include area, location and its amenities. Modelling explorations apply some regression techniques such as multiple linear regression, Lasso and Ridge regression models, support vector regression, and boosting algorithms to build a predictive model, and to pick the best performing model by performing a comparative analysis.

Predicting Future Housing Prices with Lasso Regression by Raghavendran, et al., [14] investigates the application of lasso regression in forecasting housing prices. The paper aims to showcase the efficacy of lasso regression in housing price prediction and may provide valuable insights into the optimal use of this technique in real estate valuation.

Exploring Predictive Models for House Prices by Nitish, at el., [15] delves into the intricacies of predicting house prices and significant housing characteristics within the real estate industry. Through literature research, the study identifies artificial neural networks, support vector regression, and linear regression as the most effective modelling techniques for predicting home prices.

3. SYSTEM ARCHITECTURE

The system architecture diagram outlines the sequential flow of operations in the house price prediction system. Initially, the dataset undergoes preprocessing tasks such as handling missing values and encoding categorical variables. Subsequently, the preprocessed data is divided into two subsets: a training set utilized for model training and a test set for assessing model performance. Next, individual models for Random Forest and Gradient Boosting Algorithms are constructed using the training data to predict house prices based on various features. Following model creation, the ensemble model is built by combining the predictions from both models, enhancing overall prediction accuracy. Finally, the ensemble model is deployed into production, where it serves as a predictive tool for estimating house prices. Figure 1 demonstrates the overall system architecture of House Price Prediction Model. This deployment phase involves integrating the model to monitoring its performance, and continuously updating it with new data to maintain its accuracy and relevance over time.

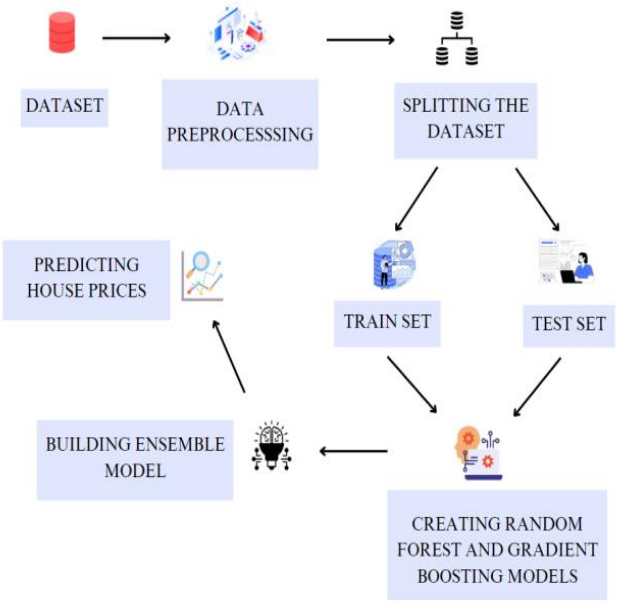

Figure – 1: System Architecture

4. METHODOLOGY

4.1 Dataset

House Price India dataset disponible on Kaggle, which contains comprehensive information on residential properties in India. This dataset offers a variety of characteristics that are valuable for analyzing and understanding the local housing market. These characteristics are as follows: Id, Date, number of bedrooms, number of bathrooms, living area, lot area, number of floors, waterfront present, number of views, grade of the house, Area of the house (excluding basement), Area of the basement, Built Year, Postal Code, Latitude, Longitude, living area renovation, lot area renovation, Number of schools nearby, Distance from the airport and Price.

4.2 Data Analyzing

Analyzing the dataset before preprocessing is an important step in better understanding of data and its properties. As part of the analysis, a correlation matrix was prepared to examine the relationship between various features. The correlation matrix has correlation coefficients between +1 and -1, indicating the correlation between two variables. Positive correlation indicates linear relationship and the negative correlation indicates non-linear relationship between the features which are independent. Analyzing the correlation matrix helps to understand the interactions between features and target variables and make informed decisions when model training.

4.3 Data Preprocessing

Data Preprocessing consists of cleaning the collected data and preparing it for training the model. Perform tasks such as handling missing values, removing outliers, normalizing numeric features, and encoding categorical variables. Specific selection criteria can be used to determine the features which are most important to estimate the house price.

4.4 Model Training

Training the model using various machine learning algorithms on previous data by using most advanced methods include Random Forests and Gradient Boosting. The training process involves fitting the model to the training data, optimizing the hyper parameters, and evaluating the performance of the model using appropriate metrics such as mean squared error or R - squared. Training sets are utilized to train the prediction models, containing abundant information to show the relationship between practical features (such as rooms, areas, square meter) and different objectives (such as house price). The model learns from the training process to make predictions.

4.5 Model Testing

Once the model is trained, it is evaluated by testing its predictive ability on test data. The model's performance is measured by comparing its predictions to actual house prices in testing. Measures such as mean square error or root mean square error can be used to measure the accuracy of forecasts. On the other hand, a separate set of data is used to measure performance and predict results. During the training phase the quality of the model predicting the house price is evaluated based on new data. Dividing a dataset into training and testing sets is usually done randomly to ensure that the two subsets have similar distributions and properties. Approximately 80% of the data to training and the remaining 20% to the testing process is allocated. The number of training rounds depends on many factors, such as the complexity of the dataset, the machine learning algorithm chosen, and the task to be performed. More training can improve the model's accuracy and improve its performance. Training each model iteratively can consume an incredible amount of time. Evaluate the relationship between the decision tree and the mean square error (MSE) of the data across 100 training sessions.

5. MACHINE LEARNING

Machine Learning (ML) is a branch of artificial intelligence (AI) where computer systems are trained to learn patterns and make decisions based on data without explicit programming instructions and accurately process large volumes of data, generating insights and predictions with minimal human intervention. ML enables organizations to streamline decision-making processes, improve productivity, and achieve better outcomes across various domains. ML includes many techniques that allow software applications to improve their performance as time progresses. It requires understanding mathematical and statistical concepts to select appropriate algorithms and training them with sufficient data to achieve accurate results. Prediction techniques leveraging machine learning algorithms across various industries to anticipate future outcomes, trends, and patterns based on historical data analysis.

Machine learning for house price prediction involves the use of computational algorithms to analyze multiple factors affecting real estate values, such as property attributes, location details, economic indicators, and past sales data. Through advanced statistical methods and mathematical models, these algorithms identify patterns in the data to create accurate predictive models. By utilizing these models, stakeholders in the real estate sector can make informed decisions regarding property investments, sales approaches, and market trends, leading to improved efficiency and optimized outcomes within the housing market. Overall, the incorporation of machine learning techniques into house price prediction systems represents a significant advancement in the field, offering enhanced accuracy, efficiency, and adaptability for stakeholders in the real estate industry.

5.1 Random Forest Regression Algorithm

Random Forest is a collection of supervised learning algorithms for classification and regression used in predictive modelling and learning. It collects the results and predictions of various decision trees and finally selects the best result, which is the class or type of the average prediction (the most common value in determining the configuration of the tree). Random Forest works by splitting the data set into two parts: the training set and the test set. More examples are then selected from the training program. Then, using the decision tree for each example split each option into two children using the best-fit split. After that the last step is repeated and all the

predictions are finally voted and the prediction with the most votes is chosen as the final result. The working of Random Forest Regression is shown in Figure 2.

The main hyper parameters in random forests are used to increase the predictive power of the model or to speed up the model. In this case, more trees can improve performance and make the prediction more stable, but unfortunately the processing time will be longer. In addition to have a minimum number of nodes to split internal nodes, using more features can improve the performance of the algorithm. Once the training step is completed, the training model can be applied to data not used for training. This method allows the forecast to be estimated and compared with expected results.

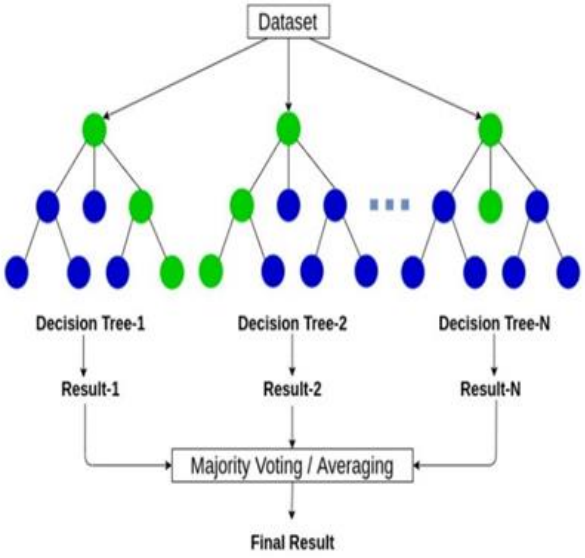

Figure – 2: Random Forest Regression

5.1.1 Algorithm: Random Forest for Regression

1. Select random K data points from the training set.
2. Build the decision trees associated with the selected data points (Subsets).
3. Choose the number N for decision trees you want to build and repeat steps 1 and 2.
4. For a new data point, find the predictions of each decision tree and assign the new data point to the average across all of the predicted values.

Random forest algorithm is based on combining decision trees and then applied to real estate data. This algorithm leverages the ability to combine multiple decision trees to make robust predictions. Its ability to capture both linear and non-linear relationships between house features makes it a suitable choice for house pricing. The prediction graph of this algorithm is shown in Chart 1.

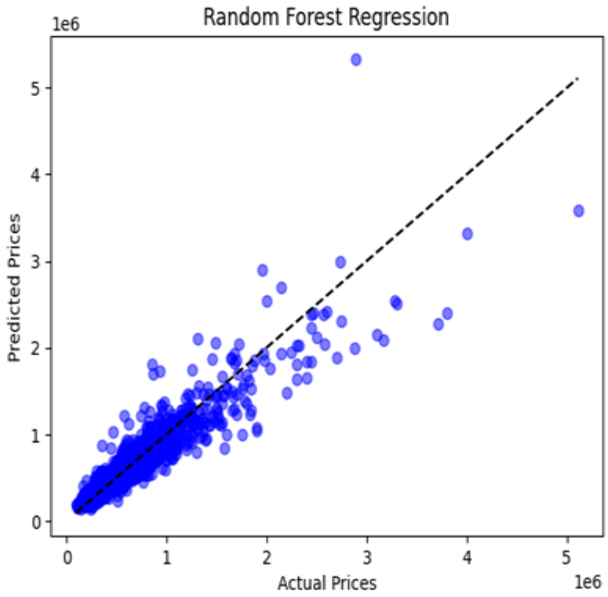

Chart - 1: Predicted Graph of Random Forest Regression

Additionally, Random Forest has a proprietary selection process that improves its ability to identify the most important aspects of house price. In a random forest, the decision tree uses the mean square error (MSE) and selects the features and values that result in the smallest MSE for each node. MSE is defined as the sum of the differences between actual and predicted house prices. Chart 2 shows the relationship between MSE and Decision Trees.

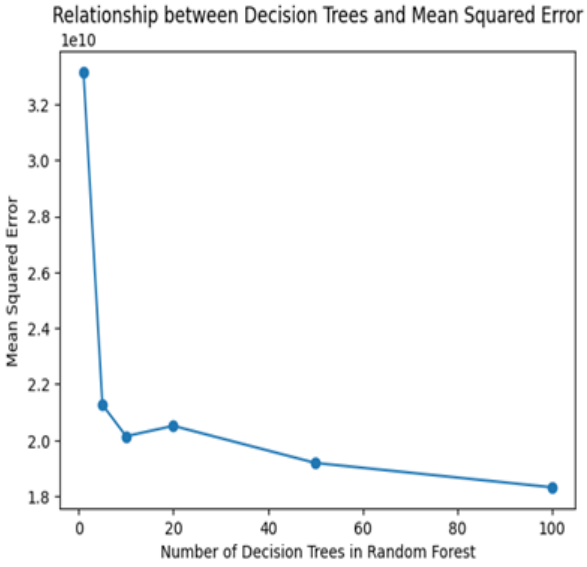

Chart - 2: Relationship between Decision Tree and Mean Squared Error

5.2 Gradient Boosting Algorithm

Gradient Boosting Regression Tree Algorithm involves learning by combining multiple regression trees (decision trees) to develop predictive models. This algorithm reduces the error of weak learning models (regressor or classifier). Weak learning models are those in which the training data has high bias, variability, and irregularity, and their results can only be considered improvements over prediction towers and are incredible. Generally, the Boosting algorithm has three components: an addictive Model, weak learners, and a loss function. The algorithm can represent non-linear relationships such as wind power curves using non-differentiable functions and can be learned through the iterative process of devices. The working of Gradient Boosting Algorithm is shown in Figure 3.

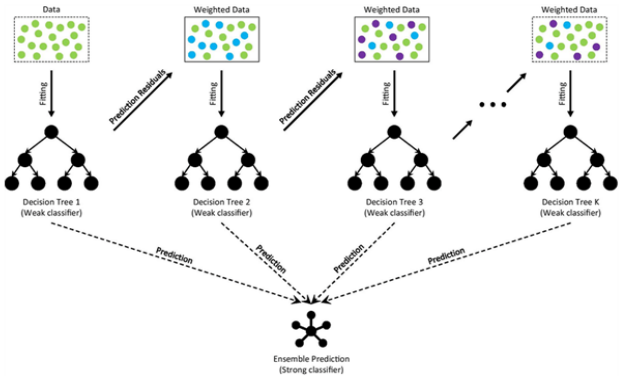

Figure - 3: Gradient Boosting Regression

Gradient Boosting Machine (GBM) works by defining the parameters of the model without the influence of gradients. This is done with the help of an iterative method where the work is eventually added to the base learner to reduce the prediction error, where the decision tree are combined by the additive model and minimizes the work by gradient descent.

5.2.1 Algorithm: Gradient Boosting for Regression

1. Consider a dataset having different data points and initialize it.
2. Now, give equal weight to each of the data points.
3. Assume this weight as an input for the model.
4. Identify the data points that are incorrectly classified.
5. Increase the weight for data points in step 4.
6. If you get appropriate output then terminate this process else follow steps 2 and 3 again.

A refined model for predicting house prices, employing the Gradient Boosting technique, enhances the accuracy of forecasts within intricate real estate environments. This method learns from data, continually enhancing predictions by analyzing past errors and refining its comprehension of the factors influencing housing prices. The prediction result of this algorithm is shown in Chart 3.

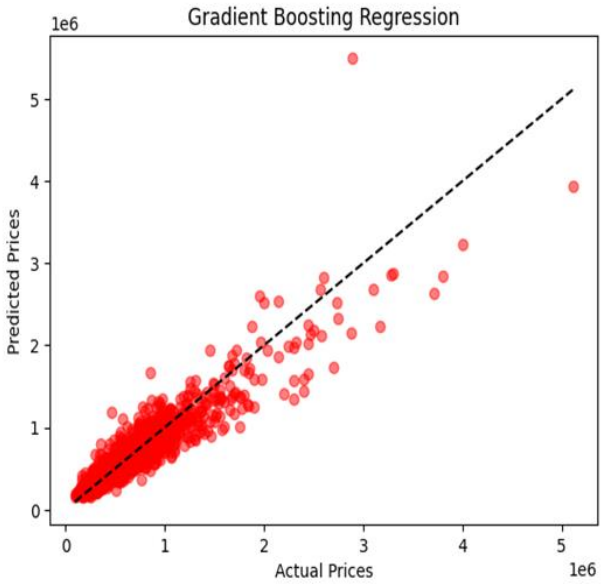

Chart - 3: Predicted Graph of Gradient Boosting Regression

The loss function for regression measures the sum of the squared differences between predicted and actual values. Chart 4 shows a graph between Deviance and Boosting Iterations.

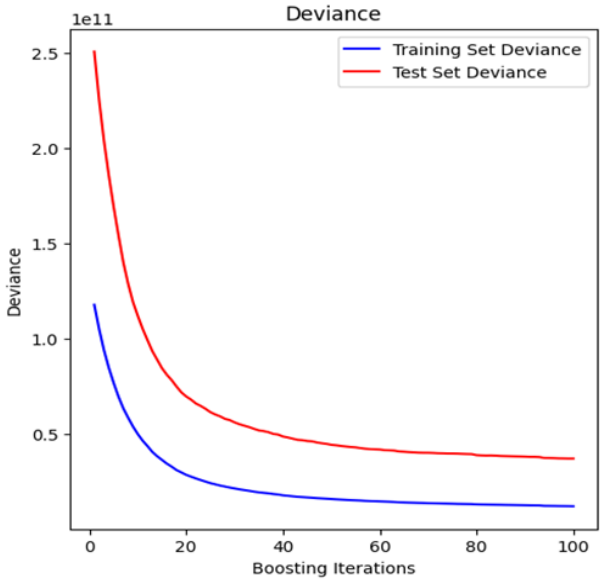

Chart - 4: A graph between the Deviance and Boosting Iterations

Finally, it provides an estimate of the house price based on the working model. This process allows informed decisions in the real estate industry and helps stakeholders to predict the house price accurately.

5.3 Ensemble Learning

Ensemble learning is a powerful machine learning technique in which multiple models are combined to improve performance. Unlike traditional approaches that rely on a single model, integration uses the intelligence of different models to produce more accurate predictions. A popular method is bagging, the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In ensemble the predictions of Random Forest (RF) and Gradient Boosting (GB) models involves aggregating their individual predictions by taking their average. By averaging their predictions in ensemble learning mitigates the risk of overfitting and variance, resulting in a more stable and reliable prediction. Chart 5 demonstrates the prediction graph of Ensemble Learning.

Ensemble learning is widely used in many fields and tasks, including classification, regression, and error detection. Its effectiveness lies in its ability to reduce overfitting, reduce bias, and improve generalization by combining predictions from multiple models. Combined methods form the basis of modern engineering practice and are often more efficient than single models.

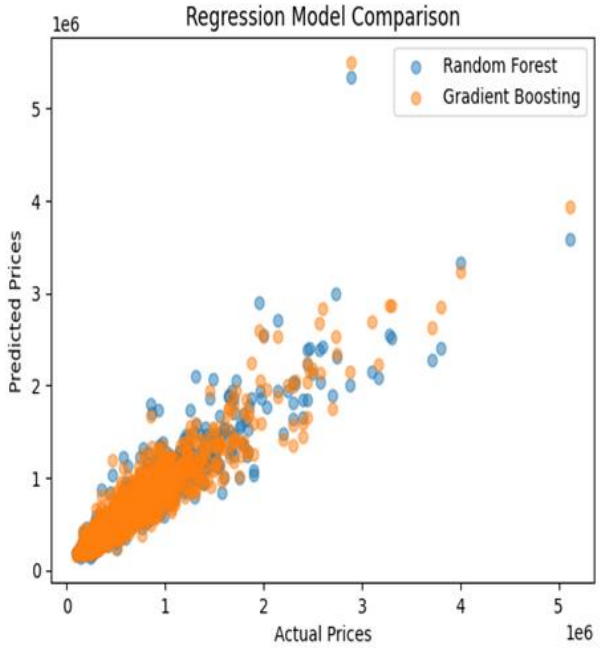

Chart - 5: Predicted Graph of Regression Models

6. EVALUATION METRICS

6.1 MSE (Mean Squared Error)

The MSE measures the average squared difference between the predicted values and the actual values. In the context of house price prediction, if y_i represents the actual price at the time i and $\hat{y}_i$ represents the predicted price at time i, then the MSE is calculated follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{..... (1)}$$

Where:

- n is the total number of observations.
- y_i is the actual house price at time i.
- $\hat{y}_i$ is the predicted house price at time i.

6.2 RMSE (Root Mean Squared Error)

The Root Mean Squared Error (RMSE) is a variant of the Mean Squared Error (MSE) that provides a measure of the average magnitude of the errors in the predictions, while still considering the scale of the data. The RMSE is calculated as the square root of the MSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{..... (2)}$$

Where:

- n is the total number of observations.
- y_i is the actual house price at time i.
- $\hat{y}_i$ is the predicted house price at time i.

6.3 MAE (Mean Absolute Error)

The Mean Absolute Error (MAE) equation is a metric used to evaluate the accuracy of a predictive model by measuring the average magnitude of errors between predicted and actual values. It is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{..... (3)}$$

Where:

- n is the number of samples or data points.
- y_i represents the actual or observed value for the i-th data point.
- $\hat{y}_i$ represents the predicted value for the i- th data point.
- $|\cdot|$ denotes the absolute value.

6.4 R^2 (R Squared Error)

R2 represents the proportion of variance in the house prices that is explained by the model. It ranges from 0 to 1, with higher values indicating a better fit. It is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{..... (4)}$$

Where:

- n is the total number of observations.
- y_i is the actual house price at time i.
- $\hat{y}_i$ is the predicted house price at time i.
- $\bar{y}_i$ is the mean of observed values of the dependent variable.

7. RESULT ANALYSIS

The combination of Random Forest (RF) and Gradient Boosting (GB) models in an ensemble achieves an accuracy of 88% in predicting house prices. This ensemble method integrates RF which handles complex interactions and outliers and GB which helps in capturing subtle patterns and boosting weaker learners. Chart 6 shows the final predicted price graph of each model. By averaging their predictions, the ensemble provides a reliable framework for house price prediction, reducing overfitting and improving generalization performance. Table 1 demonstrates the final prediction result of each model.

Table – 1: Final prediction result of each Model

Algorithm	MSE	RMSE	MAE	R^2
Random Forest	18,319,120,272.19	135,416.25	69,047.24	0.870
Gradient Boosting	18,526,489,467.85	136,135.46	77,796.18	0.869
Ensemble (RF + GB)	17,252,273,775.53	131,321.62	69,998.87	0.878

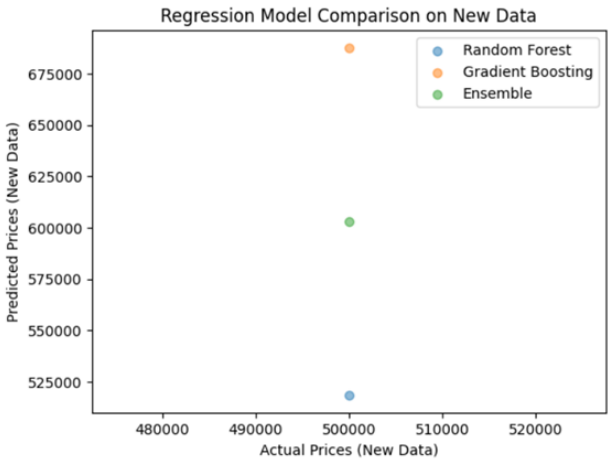

Chart - 6: Final predicted price graph of Each Models

8. CONCLUSION

House price prediction utilizes machine learning algorithms such as Random Forest and Gradient Boosting involves advanced computational techniques to analyze various factors influencing housing prices. These algorithms leverage data such as property features, location, and market trends to generate predictive models capable of estimating house prices accurately. The prediction model will allow traders or home buyers to determine the real price of a house accurately in real estate market. In summary, the impact of this model is intended to help and assist other researchers to create more accurate models that can easily and accurately predict the prices. More research on real models is needed to validate our findings. The utilization of this model enables stakeholders within the real estate sector to make well-informed decisions regarding property investments, sales, and purchases. This enhances operational efficiency and optimizes overall outcomes within the housing market by leveraging predictive insights generated through machine learning algorithms.

9. FUTURE WORK

Further exploration of data with additional features should be conducted through comprehensive feature engineering to enhance the model's predictive capabilities. It's essential to investigate advanced ensemble methods such as stacking or blending to leverage the strengths of multiple models for improved performance. Additionally, the enhancing model interpretability through techniques like feature importance analysis and SHAP values can provide insights into the factors influencing house prices. To address imbalanced data issues, consider employing sampling techniques or alternative evaluation metrics. It's crucial to develop a robust deployment strategy for the model, ensuring scalability and efficient prediction handling. Implement continuous monitoring mechanisms to track

model performance over time and detect potential issues promptly. Enhance the user interface of the application to improve user experience and usability. Lastly, incorporate a feedback loop to gather user feedback and iteratively improve the model.

REFERENCES

[1] L. Breiman, "Random forests," in Machine Learning, vol. 45, issue 1, pp. 5–32, 2001.
[2] C. C. Wang & H. Wu, "A new machine learning approach to house price estimation", in New Trends in Mathematical Sciences, vol. 6, issue 4, pp. 165–171, 2018.
[3] C. H. Raga Madhuri, G. Anuradha, M. Vani Pujitha," House Price Prediction Using Regression Techniques: A Comparative Study", 2019, IEEE.
[4] J. S. Chou, D. B. Fleshman & D. N. Truong, "Comparison of machine learning models to provide preliminary forecasts of real estate prices", *Journal of Housing and the Built Environment*, *37*(4), 2079-2114, 2022.
[5] Sifei Lu, Zengxiang Li, Zheng Qin , Xulei Yang , Rick Siow Mong Goh- "A hybrid regression technique for house prices prediction", 2017, IEEE.
[6] Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair - "Housing Price Prediction Using Machine Learning and Neural Networks", 2018, IEEE.
[7] A. S. Ravikumar, "Real estate price prediction using machine learning", 2017.
[8] John Smith, "House Price Prediction Using Machine Learning Techniques", 2018.
[9] Andrew Wang, "Predicting House Prices Using Support Vector Machines", 2021.
[10] John Smith, Emily Johnson-"A Comparative study of Regression Models for House Price Prediction"2020.
[11] Michael Brown, Sarah Davis "Spatial Analysis of House price prediction model", 2019.
[12] Christopher White, Rachel Adams, "Time Series Analysis for House Price Prediction", 2022.
[13] J. Manasa, R. Gupta, and N. Narahari, "Machine Learning based predicting house prices using regression techniques", 2020, IEEE.
[14] G. Naga Satish, Ch.V. Raghavendran, M. D. Sugnana Rao, Ch. Srinivasulu "House Price Prediction Using Machine Learning". IJITEE, 2019.
[15] M. Jagan Chauhan, D. Nitish, G. Akash, Nelli Sreevidya and Subhani Shaik, "Machine Learning Approach for House Price Prediction", Asian Journal of Research in Computer Science, vol 16, Issue 2, pp. 54-61, 2023.