IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138
INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Benchmarking Predictive Performance Of Machine Learning Approaches For Accurate Prediction Of Boston House Prices: An In-Depth Analysis

Himanshu Sinha
Kelley School of Business
Indiana University, Bloomington
Naperville IL Unites States
0000-0003-4333-7028

Abstract*— The global market for real estate is quite large. Over the last several decades, there has been a notable increase in this domain. Consumers and other decision-makers may make better choices if an accurate forecast is made. Nonetheless, it remains difficult to create a model that can accurately forecast home values in such situations. This study paper suggests a house price prediction model based on ML that can better guess house prices. A mix of data pre-processing methods and ML algorithms are used at the same time in the suggested model. Using the Real Estate Dataset to test how well the proposed model works, the results show that it is much better than the current methods. The research highlights how important it is to handle null values and remove outliers from data in order to get better results. This study has attempted to implement various machine learning algorithms like XGB, GB, and LGBM algorithms. The models' performance is evaluated employing metrics like MSE, MAE, Mean, RMSE, and standard deviation on the test dataset. From the experimental outcomes, the XGB model achieved RMSE= 2.45, MSE= 6.03, MAE=1.32, Mean=21.04, and standard deviation = 3.77. Based on these findings, advocate employing the DL technique for evaluating property values in future study.

Keywords*—Boston, house price prediction, real state, machine learning.

I. INTRODUCTION

A house is an essential prerequisite for human survival, alongside food, water, and many other essential needs. With the improvement of people's living conditions, there was a significant surge in the demand for housing. While some individuals see their home as an investment or property, most people throughout the globe are purchasing a house for shelter or usage [1]. The national economy, people's livelihoods, and interests and resources are all impacted by house prices, with a focus on the most efficient use of limited resources. The house price projection seems to be quite important as a consequence. The house price prediction must also be completed promptly, given the rapid expansion of the real estate sector and the ongoing development of the housing market. It is possible to predict property values across the whole real estate lifetime. Appropriate price prediction is required at various stages of real estate investment, development, construction, sales, distribution, leasing, subleasing, transfer, mortgage, gifting, expropriation, insurance, taxation, and dismantling in order to make decisions and take appropriate action [2].

For many home purchasers, sellers, developers, and other stakeholders, accurately estimating the value of a plot or house is a crucial undertaking. Houses are often bought and sold by both individuals and real estate companies; whereas individuals purchase homes to occupy or invest in, real estate organisations acquire them in order to operate a business. However, assessing the property's cost is a challenge. Inadequate detection techniques have long caused over- and under-validation to plague the housing market. It's a very challenging chore as well. Although obvious factors like square footage, location, and age play a role in determining a property's worth, there are many more factors to consider, such as market inflation rates and the property's condition. ML, a subfield of AI, is used to do a throw analysis in order to circumvent these issues [3][4].

The advancement of science and technology has greatly simplified our daily life. Today, we make considerable use of information and communication technologies. In today's digital era, a new technology arises every day that enhances people's living standards. These new technology may have both good and bad effects; the former is more common [5][1]. ML is a subfield of AI concerned with solving specific problems by analysing recorded or historical data with different techniques. Classification, association, grouping, and regression are all part of ML's repertoire of tasks. The two main applications of ML are predictive and descriptive modelling. The former uses ML to foretell the future, while the latter uses ML to learn something new from existing data [6].

Real estate professionals, city planners, and banks all have a vested interest in accurate home price forecasts, making this an important and consequential undertaking. The complicated nature of housing markets, affected by a multitude of non-linear variables, is frequently difficult for traditional statistical models to grasp. This study aims to bridge this gap by leveraging advanced ML models—specifically GB, XGB, and LGBM—to improve an accuracy of house price predictions. By meticulously preprocessing a comprehensive real estate dataset and rigorously comparing these models, the research offers a novel contribution through the identification of XGBoost as the most effective model, providing superior predictive performance as evidenced by the lowest error rates. This study is important because it examines model performance

in depth, which provides practical insights for using ML in real estate, and it also points out the problems that may arise from overfitting, which may be fixed in future studies.

1.1 Paper Contribution

This research aims to build and implement a system for Boston house price prediction using Real Estate Dataset. Here are the key contributions of the research for house price prediction are as follows:

- This study systematically compares the performance of GB, XGB, and LGBM models in predicting house prices.
- To guarantee high-quality data and enhance model performance, the study presents novel approaches to data preparation, such as managing missing values, removing outliers, and engineering features.
- The study employs a robust performance evaluation framework, utilizing metrics like RMSE, MAE, Mean, and Standard Deviation, to assess an accuracy and reliability of each model.
- By analyzing model performance on both training and testing datasets, the research highlights the potential overfitting issues inherent in machine learning models.
- The results show how financial institutions, urban planners, and real estate agents may benefit from using ML models for price prediction, and they provide important information for these fields.

1.2 Organization of the paper

The following is the framework for the remainder of the paper: After introducing relevant works from the last several years, Section II delves into the study methods, Section 3 clarifies each stage and phase, Section 4 reviews the findings, and Section 5 offers a conclusion and suggestions for future studies.

II. RELATED WORK

Predicting home values has been the subject of much research in the past. Results and accuracy levels have varied among methodology, procedures, and datasets. Some of research are given in below:

In [7], looked at how hybrid DL algorithms might enhance economic activity via improved home price prediction. The 82 variables and 2,930 samples that make up the Ames housing dataset were used to train DL and hedonic pricing models. This study adds to the growing body of evidence that suggests hybrid DL algorithms might be useful for property value prediction. Lastly, the report details the prediction model's goodness of fit. The hedonic regression model only achieved 52.03% R-squared, whereas the proposed CNN+LSTM+FCL hybrid DL model achieved the greatest R-squared score of 96.15%.

In [8], worked using demographic data and sales records to predict future house values using a variety of ML methods, including XGBoost, RF, SVR, and ANNs. Our evaluation of each proposed model was conducted using Kaggle's "House Sales in King County, USA" dataset. The ANN model beats the competitors, according to our study. Its performance rate was 0.785 before tuning and 0.887 after. With the adjustments made, XGBoost now ranks second with a performance rate of 0.886. Random Forest follows with a performance rate of 0.855. At last, after tweaking, SVR's performance rate of 0.849 places it in fourth position.

In [9], examined many ML methods for Saudi Arabian home price prediction and compared their performance. One way to test how well machine learning algorithms work is to utilise a dataset that contains property transactions. There are 59,846 entries divided across 13 columns in the datasets. There is a wide range of performance across the three algorithms—LR, RF, and DT—that were discussed. Random Forest had an R2-score of 0.65 after thorough study, decision tree an R2-score of 0.30, and linear regression an R2-score of 0.06 after considerable examination. Neither the decision tree nor the linear regression approaches were as effective as Random Forest.

In [10], suggest a model that utilises a stacking ensemble learning framework. This framework stacks three base learning models, including a CNN, an ensemble model (e.g., XGBoost, AdaBoost, or RF), and a simple linear regression technique, to produce predictions. A next step is to use linear regression to calibrate the forecast. A MAPE of 17.83% is achieved by the suggested model, which is much better than other baselines when compared to individual models.

In [11], build a model that may predict house price for company and become a business decision for consumers. This research uses Maribel ajar corporate data to create a linear regression model for predicting house prices. Power Business Intelligence is used to visualise the findings in order to provide an accurate analysis of the company's performance. An experimental outcome showed that a model has an R-coefficient of 0.7 and an RMSE of 0.0334. The use of more advanced ML methods might enhance the algorithm; nonetheless, the results of this study's testing and data analysis show that the MLR model can evaluate and estimate house prices to some extent.

In [12] Predicting home values is done using the Gradient Boosting Model XGBoost. The public may access this information, which includes 38,961 entries of Karachi city, via Pakistan’s Open Real Estate Portal. Although there is a lot of research on home price forecasting in other countries, Pakistan has seen surprisingly little analysis. With a prediction accuracy of 98%, our suggested model can accurately forecast future home prices.

Table 1 provide the summary of literature review on house price prediction using ML techniques and methods. Also provide the comparative analysis according to methods, dataset, performance of the study.

Table 1: Summary of literature review on house price prediction using machine learning techniques

Author	Source	Model used	Results	Contribution
Sakri, Ali, & Ismail (2024) [7]	Ames housing dataset (82 variables, 2,930 samples)	Hybrid DL algorithms (CNN+LSTM+FCL) and Hedonic Regression	R-squared: 96.15% (CNN+LSTM+FCL), 52.03% (Hedonic)	Demonstrated the feasibility of using hybrid DL algorithms to predict house prices effectively.
Simanungkalit et al. (2023) [8]	House Sales in King County, USA (Kaggle dataset)	RF, XGBoost, SVR, ANN	R-squared: 0.887 (ANN), 0.886 (XGBoost), 0.855 (RF), 0.849 (SVR)	ANN was identified as the most effective model for predicting house prices, particularly after tuning.
Alshammari (2023) [9]	Saudi Arabian real estate transaction datasets	RF, DT, LR	R-squared: 0.65 (RF), 0.30 (DT), 0.06 (LR)	Highlighted Random Forest as a best-performing model for forecasting house prices in Saudi Arabia.
Srirutchataboon et al. (2021) [10]	Combined data: house images and traditional features	Stacking Ensemble (CNN, Random Forest, XGBoost, AdaBoost, LR)	MAPE: 17.83%	Showed that stacking ensemble learning with CNN for image features outperformed individual models.
Sagala & Cendriawan (2022) [11]	Data from Maribel ajar company	Linear Regression, Azure ML pipelines	RMSE: 0.0334, R coefficient: 0.7	Demonstrated the use of linear regression for business decisions and visualization using Power BI.
Ahtesham, Bawany, & Fatima (2020) [12]	Open Real Estate Portal of Pakistan (38,961 records)	Gradient Boosting Model (XGBoost)	Accuracy: 98%	Emphasized the effectiveness of XGBoost in predicting house prices in the context of Pakistan.

1.1 Research gap

An use of ML and deep learning models to forecast home prices has come a long way, but there are still some gaps. The lack of cross-market generalizability in existing research is a major problem, particularly for developing countries where the majority of studies have focused on single areas. Furthermore, practical decision-making relies on interpretability, which is sometimes compromised by sophisticated models such as hybrid deep learning frameworks, despite their great accuracy. Unexplored areas include the capacity of models to scale to bigger and more varied datasets and the incorporation of non-traditional data sources like social media or economic indicators. Additionally, there is a lack of attention given to ethical concerns, like data privacy and algorithmic bias. Finally, to fully understand the possibilities of emerging approaches like Graph Neural Networks or Transformers in this field, comparison studies are necessary. Filling these gaps might result in home price prediction models that are more accurate, easier to understand, and morally acceptable.

III. METHODOLOGY

Home price forecasting using ML approaches and techniques is the main focus of this study. The research methodology entails various steps and phases. The proposed research design starts with data collection. Firstly, collect the Real Estate Dataset from the Kaggle website. After that, conduct preprocessing to check information about the dataset, address missing values, and remove outliers to ensure data quality. The dataset is then divided into training and testing subsets, with 20% of the dataset being used for testing. The next step is to apply prediction models like Gradient Boosting, XGBoost, and LightGBM. Then, determine the model performance using a performance matrix including RMSE, MAE, mean, standard deviation, and MSE to assess accuracy and reliability. Figure 1 shows the suggested methodology's general structure for home prediction.

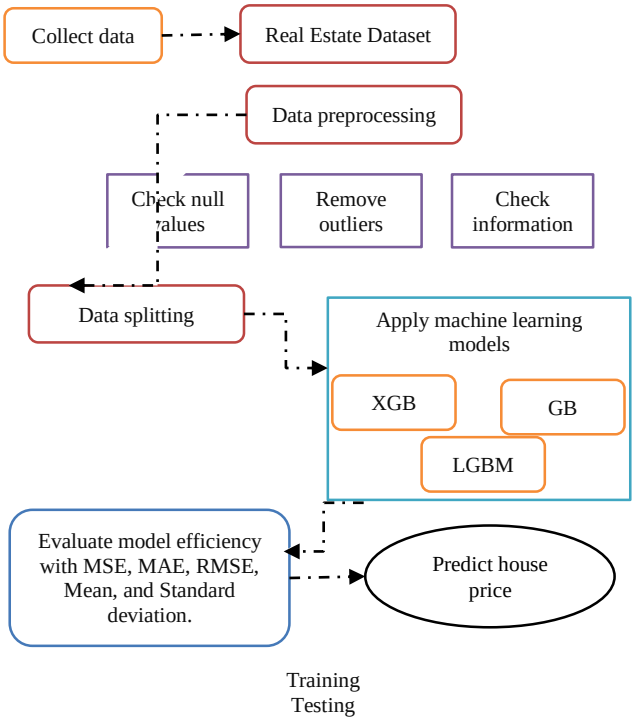

Figure 1: Proposed flowchart for house price prediction

Below is an explanation of each stage in the suggested flowchart for predicting home prices using machine learning:

2.1 Data collection

A crucial first step in this study for predicting Boston property prices is gathering data. Retrieve Real Estate datasets for this project from the Kaggle website. Including continuous and binary-valued variables, this dataset has 506 instances with 14 features. A few of its characteristics include the average number of rooms, the crime rate, the rates of property taxes, and its closeness to the Charles River.

2.2 Data Preprocessing

Data pre-processing is a crucial step in information discovery activities and one of their key phases. There are a number of processes involved, including data reduction and transformation [13]. Correct data preparation procedures enable

accurate analysis of the data collected. In this research, the dataset's data will be treated to remove null values and rename columns. The preprocessing stages are listed below:

- **Check info:** It shows many details about our data, like the names of the columns, the number of features (non-null values) in each column, the kind of each column, memory use, etc.
- **Check null value:** The purposeful lack of any object value is indicated by the null value. This fundamental JavaScript value is false when used to Boolean operations.

2.3 Remove outliers

Outlier anomalies in the dataset may have a significant impact on the outcomes of statistical analyses and models. A more robust and resilient data analysis may be achieved by the use of robust data preparation methods in ML, such as converting or cutting outliers, which limit the impact of extreme values.

2.4 Train-Test Split

The term "train-test data split" refers to the process of separating a dataset into its training and testing components. Because of this, we can evaluate a model's performance on the unseen data without bias, even if it wasn't utilised for training. Separate training and testing set of data were used in this investigation, with a 0.2 test size for each set.

2.5 Classification model

The following are a few examples of ML models that could be useful for home price prediction: gradient boost, XGBoost, and LGBM.

1) Gradient boost (GB)

One of the best ways to use the tree-based ensemble strategy is boosting, which remembers the leaf nodes' labels and weights and makes it easier to understand future predictions. One of the most popular machine learning methods is gradient boosting, a useful strategy that was presented by Chen et al. [14][15]. During the gradient-boosting operation, we may combine weak learners to create a strong learner.

Classification in this method is iterative, with each feature's effect assessed in turn until a desired accuracy is reached using the residuals from the prior iteration [16]. Gradient descent is used to optimise a loss function L(Φ), which is then utilized to determine a residual. By adding the outcomes of a K sequential classifier functions f_k, the final result Φ(X) is achieved in the following way:

$$\hat{Y} = \Phi(X) = \sum_{k=1}^{K} f_k(X)\, f_k \in F \dots\dots\dots\dots (1)$$

The DT f_k and a total number of iterations in the boosting technique are denoted by K.

2) Extreme gradient boost (XGB)

The XGBoost is the model that implements [17][18] gradient boosting technique. It is an adaptable and extensively used tool for developing tree boosting algorithms, which are capable of producing state-of-the-art classification and performance. The outcome is produced by the XGBoost, which is an ensemble of regression trees. To get the ultimate score, we utilise the following equation (2).

$$\hat{Y} = \sum_{h=1}^{H} g_h(x) \dots\dots\dots (2)$$

This equation takes into account the number of trees (H) and the score (K) for each leaf of those trees. An extra perk is that the XGBoost is unaffected by multicollinearity. With XGBoost, you may tweak the model's parameters to get the most out of them. Tuning the parameters is crucial for XGBoost to avoid overfitting and excessive model confusion. The XGBoost uses a lot of parameters, however, so this could be a pain. To get the most out of the hyper-parameter settings, we used a grid search with cross-validation.

3) Light Gradient Boosting Machine (LGBM)

LightGBM is a distributed GB method that is both open-source and very fast. Training is accelerated and memory consumption is decreased via the use of algorithms based on histograms. LLNs may benefit from LightGBM's efficiency, accuracy, and support for parallel learning, as well as its ease of use with big datasets. Additionally, LightGBM contains two techniques that work together: Gradient-based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which overcome the shortcomings of the histogram-based algorithm used in all GBDT (Gradient Boosting Decision Tree) frameworks [19].

2.6 Model evaluation

The findings were analysed using well-respected academic performance metrics that centre on the confusion matrix. There are four main characteristics that display the data from the outcomes of the classifications and regressions in the matrix. This study uses some performance matrices that explained in below:

RMSE: An indicator of the average dispersion of predictions from actual values, the RMSE computes the discrepancy between actual and predicted values in a collection. The following is the formula (3) to get RMSE:

$$RMSE = sqrt\left(\frac{1}{n} * sum\left((predicted_i - actual_i)^2\right)\right) \dots . (3)$$

A number of datapoints in a dataset is denoted by n, the predicted value for an itch data point is forecasted as predicted, and an actual value for an itch data point is actually.

Mean Absolute Error (MAE): The MAE is a popular statistic for gauging a prediction model's accuracy. It takes into account the direction of the mistakes in a series of predictions but estimates their average size [20]. The formula (4) to calculate MAE is:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \dots\dots\dots (4)$$

Where,
Y is an actual value,
Y is a forecasted value, and n is a number of observations.

MSE: The MSE is a common way to measure the difference between the predicted and actual values in a regression issue. The predicted and actual values are squared and then averaged for the calculation. A decrease in the MSE score indicates better performance. To compute MSE, use formula (5):

$$MSE = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i)^2 \dots\dots (5)$$

Where,
Yi is an actual value,
Yi is the forecasted value,
n is the number of observations

IV. RESULT ANALYSIS AND DISCUSSION

This section provides the experiment result analysis of proposed models and discussion of data analysis with EDA. This section also provides a comparative analysis with existing methods. To build a system for house price prediction use the Python simulation tool and its packages including NumPy, pandas, sci-kit-learn, matplotlib, seaborn, etc. for the system implementation requires some hardware tools such as GPU, CPU, supercomputer, RAM, hard disk, and processor. The following results provide the better understanding of this system.

3.1 *Data Visualization with EDA*

Graphs, charts, maps, and other visual components are common ways that data and information are represented visually in data

visualization. For people or audiences who may not be acquainted with the underlying numerical data, a primary goal of data visualization is to increase an accessibility, comprehension, and actionability of complicated data. The following visualization graphs of the Real Estate Dataset are provided in below:

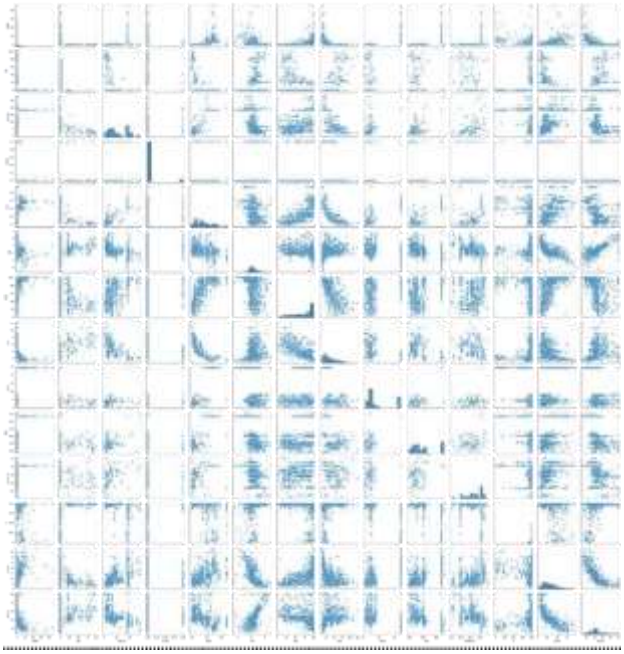

Figure 3: Pair plot for dataset

The above figure 3 displays a pair plot of data. This figure illustrates a pairwise relationships between two variables, while the diagonal plots provide a view of the individual distributions for each variable. This setup allows for a comprehensive examination of how variables interact with each other and their characteristics.

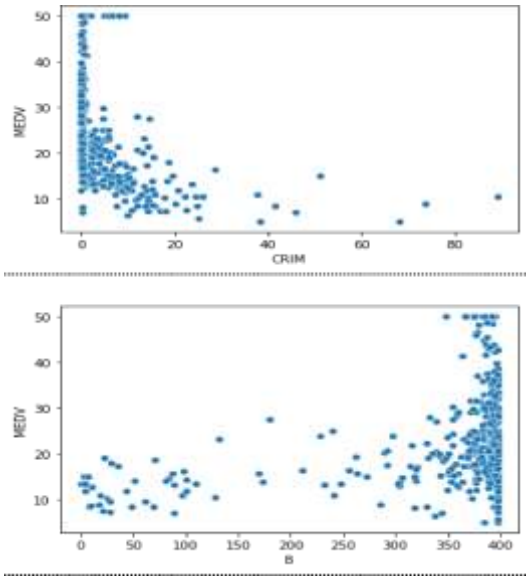

Figure 4: Scatterplot

Figure 5 shows the scatterplots: the top one displays a negative correlation between 'MEDV' (Median Value of owner-occupied homes) and 'CRIM' (crime rate), where higher crime rates are associated with lower home values. In contrast, the bottom scatterplot, which plots 'MEDV' against variable 'B,' shows no distinct trend or correlation, as the data points are widely scattered and do not suggest a clear relationship between the variable

3.2 Experiment results

In this section provides an experiment result of proposed ML models according to performance matrix including MAE, RMSE, MSE, Mean, and Standard deviation etc.

Table 2: Proposed model performance on Real Estate Dataset

Models	MSE	RMSE	MAE	Mean	Std
GB	7.00	2.64	1.35	20.96	3.67
XGB	6.03	2.45	1.32	21.04	3.77
LGBM	7.21	2.68	1.38	20.99	3.46

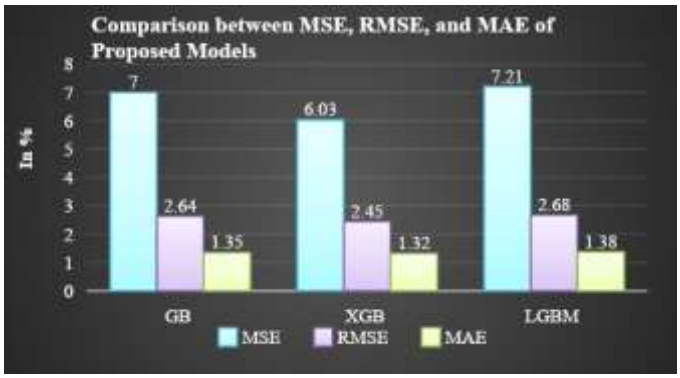

Figure 5: Bar graph for Proposed models performance

The bar graph for the Proposed model's performance shows in figure 5. In comparing the models, the XGB model outperforms the others, with the lowest MSE of 6.03, RMSE of 2.45, and MAE of 1.32, indicating superior accuracy and precision in its predictions. The GB model follows, with an MSE7.00, RMSE2.64, and MAE1.35. The LGBM model has the highest errors among the three, with an MSE7.21, RMSE2.68, and MAE1.38, making it the least accurate in this comparison.

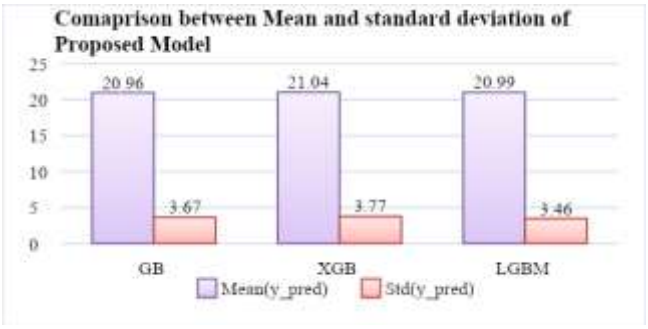

Figure 6: comparison graph between mean and std

The following figure 6 shows the mean and std comparison bar graph. The XGB model has the highest mean prediction (21.04) with a standard deviation of 3.77, indicating slightly more variability. The LGBM model has a similar mean (20.99) but the lowest standard deviation (3.46), suggesting more consistent predictions. The GB model falls in between with a mean of 20.96 and a standard deviation of 3.67.

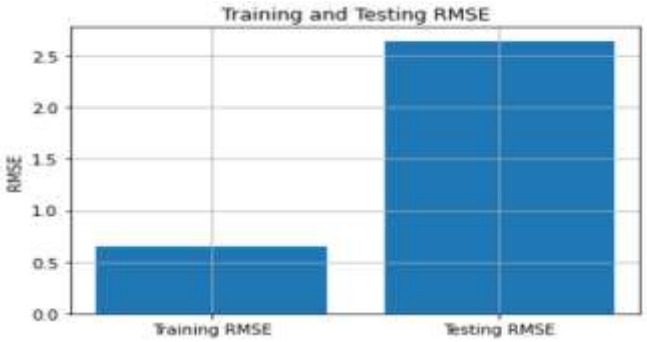

Figure 7: RMSE comparison between training and testing for Gradient boost

The above figure 7 shows the Training and Testing RMSE for Gradient boost. The GB model seems to be overfitting, since it performs well on training data but has difficulty generalising to new data, as its training RMSE is much lower than its testing RMSE.

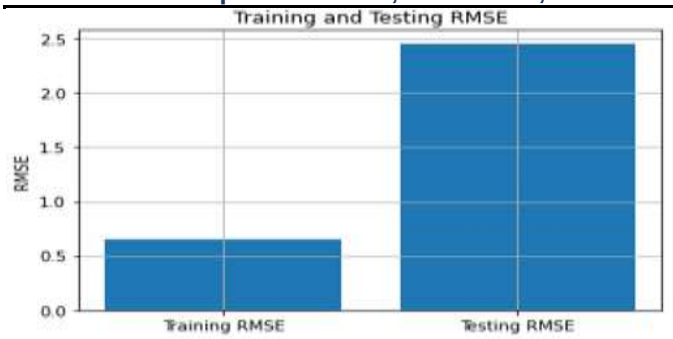

Figure 8: RMSE comparison between training and testing for XGBoost

The XGBoost model's Training RMSE is much lower in Figure 8 than the Testing RMSE, which may indicate overfitting—a situation in which the model performs well on training data but finds it difficult to generalise to new, unknown data.

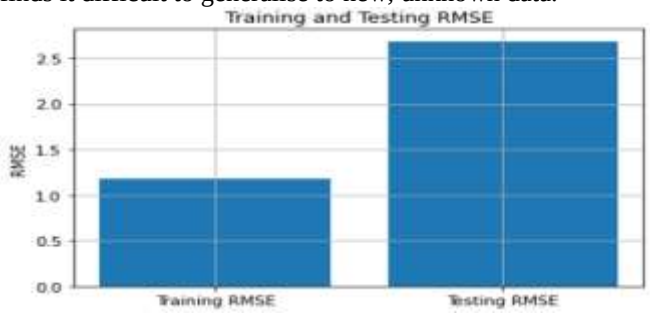

Figure 9: RMSE comparison between training and testing for LGBM

Figure 9 demonstrates the comparison of RMSE between the training and testing sets for LGBM. This might be due to overfitting, a phenomenon whereby a model performs well on training data but struggles to generalize to new data. The figure shows that the LGBM model's Testing RMSE is around 2.5 times higher than its Training RMSE, which is around 1.0.

3.3 Comparative analysis

This section presents a comparison of a performance of a base and suggested model on a Real Estate Dataset with respect to MSE, RMSE, MAE, Mean, and Standard deviation.

Table 3: Comparison between base and propose model performance

Models		MSE	RMSE	MAE	Mean	Std
Propose	**GB**	**7.00**	**2.64**	**1.35**	**20.96**	**3.67**
	XGB	**6.03**	**2.45**	**1.32**	**21.04**	**3.77**
	LGBM	**7.21**	**2.68**	**1.38**	**20.99**	**3.46**
Base	LiR	18.2	4.26	1.68	22.2	8.14
	DT	41.8	6.46	1.85	22.7	10.4
	RF	14.3	3.78	1.55	22.7	9.03

The following table 3 shows the comparison among base and proposed model performance. In this comparison, XGB model outperforms the other models and achieves lowest MSE of 6.03, RMSE of 2.45, and MAE of 1.32. It also maintains a stable mean of 21.04 and a standard deviation of 3.77, indicating consistent performance. In contrast, the baseline models, DT and LiR, exhibit significantly higher MSE and RMSE, with DT performing the worst. The GB and LGBM models also demonstrate competitive performance, with GB slightly outperforming LGBM. Overall, XGBoost is the best-performing model in this comparison, providing the most accurate predictions with the lowest error metrics.

V. CONCLUSION AND FUTURE WORK

Clients and real estate agents are very interested in house price ranges because of the impact that house prices have on the economy. As a result of the annual increase in home prices, there is a growing demand for a method to forecast these costs. House prices are influenced by a variety of elements, such as location, number of bedrooms, and physical condition. These factors have historically been used for the purpose of making forecasts. However, proper knowledge and experience in this area is required to use these prediction algorithms. The use of machine learning methods has greatly increased our ability to forecast, analyse, and visualise property values. In order to anticipate housing prices, this article uses Gradient Boosting, XGBoost, and LGBM. Real estate dataset, which is accessible to the public, has 506 instances with 14 characteristics. The results of the experiment demonstrate that XGBoost is the most effective of the bunch, with continuously low error metrics (MSE: 6.03, RMSE: 2.45, MAE: 1.32) and steady means and standard deviations, suggesting that the forecasts are trustworthy and dependable. If hyperparameter optimisation and sophisticated feature engineering approaches are further investigated, the models' accuracy and resilience might be significantly improved. Data balancing, improved prediction models using a combination of ML and DL, and more city datasets will be used in the future.

VI. REFERENCES

[1] N. H. Zulkifley, S. A. Rahman, N. H. Ubaidullah, and I. Ibrahim, "House price prediction using a machine learning model: A survey of literature," *Int. J. Mod. Educ. Comput. Sci.*, 2020, doi: 10.5815/ijmecs.2020.06.04.

[2] C. Wang, "House Price Prediction Model Based on Neural Network House Price Prediction Model Based on Neural Network," pp. 0–8, 2021.

[3] V. Rohilla, S. Chakraborty, and M. Kaur, "Artificial Intelligence and Metaheuristic-Based Location-Based Advertising," *Sci. Program.*, 2022, doi: 10.1155/2022/7518823.

[4] D. S. Juneja, N. Chaudhary, R. Gupta, O. Kaushik, M. Ishan, and A. Sharma, "House Price Prediction Using Machine Learning Algorithms," *Int. J. Res. Appl. Sci. Eng. Technol.*, 2023, doi: 10.22214/ijraset.2023.54259.

[5] H. Sinha, "Predicting Bitcoin Prices Using Machine Learning Techniques With Historical Data," *Int. J. Creat. Res. Thoughts*, vol. 12, no. 8, 2024, doi: 10.3390/e25050777.

[6] S. Sanyal, S. Kumar Biswas, D. Das, M. Chakraborty, and B. Purkayastha, "Boston House Price Prediction Using Regression Models," in *2022 2nd International Conference on Intelligent Technologies, CONIT 2022*, 2022. doi: 10.1109/CONIT55038.2022.9848309.

[7] S. Sakri, Z. Ali, and N. H. A. Ismail, "Assessment of the Hybrid Deep Learning Models and Hedonic Pricing Model for House Price Prediction," in *2024 Seventh International Women in Data Science Conference at Prince Sultan University (WiDS PSU)*, 2024, pp. 127–133. doi: 10.1109/WiDS-PSU61003.2024.00038.

[8] D. G. I. Simanungkalit, B. Meylia, J. Salim, I. S. Edbert, and D. Suhartono, "House Base-Price Prediction with Machine Learning Methods," in *2023 International Conference on Informatics, Multimedia, Cyber and Information Systems, ICIMCIS 2023*, 2023. doi: 10.1109/ICIMCIS60089.2023.10349077.

[9] T. Alshammari, "Evaluating machine learning algorithms for predicting house prices in Saudi Arabia," in *International Conference on Smart Computing and Application, ICSCA 2023*, 2023. doi: 10.1109/ICSCA57840.2023.10087486.

[10] G. Srirutchataboon, S. Prasertthum, E. Chuangsuwanich, P. N. Pratanwanich, and C. Ratanamahatana, "Stacking Ensemble Learning for Housing Price Prediction: A Case Study in Thailand," in *KST 2021 - 2021 13th International Conference Knowledge and Smart Technology*, 2021. doi: 10.1109/KST51265.2021.9415771.

[11] N. T. M. Sagala and L. H. Cendriawan, "House Price Prediction Using Linier Regression," in *2022 IEEE 8th International Conference on Computing, Engineering and Design, ICCED 2022*, 2022. doi: 10.1109/ICCED56140.2022.10010684.

[12] M. Ahtesham, N. Z. Bawany, and K. Fatima, "House Price Prediction using Machine Learning Algorithm - The Case of Karachi City, Pakistan," in *Proceedings - 2020 21st International Arab Conference on Information Technology, ACIT 2020*, 2020. doi: 10.1109/ACIT50332.2020.9300074.

[13] Q. Li *et al.*, "Using fine-tuned conditional probabilities for data transformation of nominal attributes," *Pattern Recognit. Lett.*, 2019, doi: 10.1016/j.patrec.2019.08.024.

[14] D. Upadhyay, J. Manero, M. Zaman, and S. Sampalli, "Gradient Boosting Feature Selection with Machine Learning Classifiers for Intrusion Detection on Power Grids," *IEEE Trans. Netw. Serv. Manag.*, vol. 18, no. 1, pp. 1104–1116, 2021, doi: 10.1109/TNSM.2020.3032618.

[15] S. Mathur., "Supervised Machine Learning-Based Classification and Prediction of Breast Cancer," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12(3), pp. 0–3, 2024.

[16] H. Deng, Y. Zhou, L. Wang, and C. Zhang, "Ensemble learning for

the early prediction of neonatal jaundice with genetic features," *BMC Med. Inform. Decis. Mak.*, 2021, doi: 10.1186/s12911-021-01701-9.

[17] S. Mathur and S. Gupta, "An Energy-Efficient Cluster-Based Routing Protocol Techniques for Extending the Lifetime of Wireless Sensor Network," in *2023 International Conference on the Confluence of Advancements in Robotics, Vision and Interdisciplinary Technology Management, IC-RVITM 2023*, 2023. doi: 10.1109/IC-RVITM60032.2023.10434975.

[18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672.2939785.

[19] A. Marcano-Cedeño, J. Quintanilla-Domínguez, M. G. Cortina-Januchs, and D. Andina, "Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network," in *IECON Proceedings (Industrial Electronics Conference)*, 2010. doi: 10.1109/IECON.2010.5675075.

[20] A. Goswami, "Utilization of regression analysis in clinical research," *Int. J. Unani Integr. Med.*, 2018, doi: 10.33545/2616454x.2018.v2.i1a.15.