MARKET DATA ANALYSIS AND APPLICATION FOR ASSETS COMPUTATION AND RECOMMENDATION

Prof. Dharamvir[1], Gayathri S[2] , Harish M[3], Dilip Kumar[4], Daniya Kouser[5], G Bharath[6]

Article History: Received: 15.08.2022 **Revised:** 16.10.2022 **Accepted:** 22.12.2022

Abstract :

There are numerous housing options available in today's society, so anyone can choose an affordable home that is close to nature and equipped with all the necessary amenities. However, the rising cost of housing has made it difficult to construct new housing types economically for the public benefit. Rather than estimating a single number, it is sometimes more useful and enticing to forecast a range of property value declines. Due to the difficulty in categorizing products, it may be difficult to estimate pricing. Academics frequently use the House Price Index (HPI) query method to attempt accurate forecasts of future house price fluctuations by monitoring average price changes across multiple acquisitions or refinancing operations involving identical properties. The fact that real estate price trends are affected by multiple variables, such as population density and geographic location, adds to the complexity of this issue. In this study, we use four algorithms to estimate and propose housing prices: linear regression, a decision tree, a random forest regressor, and a gradient boost regressor. The purpose of these steps is to develop a dependable machine-learning model for use in predictive analytics and data classification. We have also contrasted the accuracy of each individual's predictions. The research utilizes the Bangalore dataset, which contains more than 13,000 instances. Due to the high quality of its results, Gradient Boosting is an efficient method for recommending secure real estate investments. The Gradient Boost Regressor, the most accurate of these options, was chosen as the primary model.

Keywords : Price Indexing , System Automation , Data fragmentation , Machine Learning

[1]Assistant Professor , Dept. of MCA, The Oxford College Of Engineering, Bengaluru, Karnataka, India- 560068
[2,3,4,5,6]MCA Final Year, Dept. of MCA, The Oxford College Of Engineering, Bengaluru, Karnataka, India- 560068

Email: [1]dhiruniit@gmail.com, [2]gayathrismca2023@gmail.com

DOI: 10.31838/ecb/2022.11.12.47

1. Introduction

The dynamic nature of the housing market poses significant obstacles for homebuyers today. They must consider their lifestyle preferences, their financial circumstance, and the significance of a safe, pleasant, and readily accessible residence. However, the rising cost of housing has highlighted the difficulty of designing housing that is both affordable and secure for its inhabitants. It is more difficult than ever to generate an accurate estimate of housing values due to the numerous factors in play, including supply and demand, population growth, and urban regeneration. As it is difficult to accurately predict pricing expectations over time due to the unpredictability of the market, researchers have conducted extensive studies to comprehend how these various factors interact with housing prices in order to develop models capable of capturing market behaviour more precisely. The House Price Index (HPI) is a popular instrument for monitoring how the market value of a property responds to changes in variables such as the number of prospective buyers, the number of available sellers, and mortgage interest rates.
Due to the incapacity of a single machine learning model to generalise effectively, it is challenging to make accurate predictions based on house price data. Ensemble learning models combine several inferior learners, either homogeneous or heterogeneous base learners, into a single robust learner to enhance the generalizability of the prediction model.

The objective of this study is to develop a machine learning model that can be used to forecast and categorise property prices and then provide price-related recommendations. We examine four prominent algorithms (Linear Regression, Decision Tree, Random Forest Regressor, and Gradient Boost Regressor) that can help us arrive at our destination. The pervasive use of these algorithms in the field of home price research is a direct result of the significant advantages they offer in recognising and predicting price trends.

This study aims to assess the ability of various algorithms to predict future home prices in order to identify the most optimal model for this endeavour. Based on our findings, the Gradient Boost Regressor algorithm is the most effective method for predicting and recommending future property prices. We evaluate the efficacy and precision of these four techniques in comparison to three other cutting-edge models, including XGBoost, Lasso Regression, and Neural Networks.

This study focuses on only four algorithms for predicting property prices, with an emphasis on Gradient Boosting and its function in producing accurate price calculations and reliable recommendations. We work diligently to provide people with an invaluable resource for making informed decisions about where to reside, how to manage their finances, and how to promote a healthy community.

This article provides a comprehensive evaluation of the approaches taken, including the data analysis techniques, the criteria for determining whether or not the trials were successful, and the advantages and disadvantages of the algorithms employed. Despite the volatile nature of the housing market, our ultimate goal is to develop trustworthy machine learning models that can assist individuals in their search for reasonably priced housing.

Literature Review
Real estate research has focused extensively on the forecasting of property prices and the development of effective machine learning models for calculation and recommendation. Forecasting real estate prices and providing pertinent guidance can be difficult; consequently, a number of algorithms and methods have been investigated to assist.

Anand G. Rawool et al. made their housing price projections using 2021 survey data. Among the Machine Learning techniques used to construct a predictive model are Linear Regression, Decision Tree, and Random Forest. From data collection and cleansing to data analysis and model development, they followed a systematic, sequential process. Each completed model is then evaluated, and the resulting information is saved to a text file. Compared to the original training data, these Random Forests produce superior results.Random Forest yielded the highest accuracy, approximately 87% [1], when compared to other methods.

P.Durganjali proposed estimating the future sale price of a property using classification algorithms. In this study, we predict prospective property resale values using a variety of classification algorithms, including Leaner regression, Decision Tree, K-Means, and Random Forest. A home's price can vary depending on its physical characteristics, location, and economic climate. Using the root-mean-squared error (RMSE) as a performance matrix, we apply these methods to a variety of datasets and identify the model with the highest predictive accuracy.
Sifei Lu has introduced a novel regression method for valuing real estate. This study examines a novel technique for feature engineering in light of the limited dataset and data characteristics. The "House Price: Advanced Regression Techniques" Kaggle competition recently utilised this method as its principal premise. This study's primary objective is to calculate a reasonable pricing for consumers by considering their income levels and individual preferences[6].

CH.Raga Madhuri proposed a comparative analysis of regression techniques for estimating housing prices. This study compares and contrasts a number of well-known machine learning techniques, including multiple linear regression, ridge, LASSO, Elastic Net Gradient boosting, and Ada Boost Regression, to determine their similarities and differences. We have investigated these methodologies using a particular dataset[7].This information should be beneficial for real estate sellers establishing competitive prices and homebuyers determining the optimal periods to purchase. When estimating the ultimate price of a property, our team considers a variety of factors, including the property's model analysis, physical condition, floor plan requirements, and location.

Andrey Viktorovich Parastovich proposed Regression Methods in Machine Learning for Predicting Home Sales Prices. This research seeks to develop a method for doing so. Our method integrates conventional machine learning techniques with innovative concepts such as the residual regressor, logit transform, and neural network machine. This solution won the machine learning competition "House Prices: Advanced Regression Techniques" on the Kaggle platform. The goal was to estimate how much a property would fetch on the market based on its square footage and year of construction[8].

Adyan Nur Alfiyatin proposed employing particle swarm optimisation and multivariate regression analyses to forecast future property prices. This study contrasted the predictive power of linear regression and particle swarm optimisation in predicting future housing market values quantitatively. To examine the relationship between past and present prices, time series data are analysed quantitatively.In addition, they employ linear regression with hedonic pricing as a secondary method. The least-squares method is a reliable but time-consuming method for calculating coefficients that is utilised by linear regression. Utilising particle swarm optimization[9], the optimal coefficient values were determined.

Fan et al. [10] analysed the correlation between various property characteristics and their resale values using a decision tree. Examining the relationship between house prices and characterising characteristics, we employ a hedonic-based regression strategy in this study. Ong et al. [11] and Berry et al. [12] have also utilised hedonic-based regression to estimate the value of a home based on its most distinguishing characteristics.
By comparing the performance of several home value forecasting algorithms, we hope to contribute new knowledge to the existing corpus of literature. Such techniques include Linear Regression, Decision Trees, Random Forests, and Gradient Boosting. College degrees are beneficial because they equip graduates with the knowledge and experience necessary to find and maintain satisfying careers. The OECD discovered that college graduates earn 56% more than their non-graduate counterparts, proving that college education is a wise investment.

Our primary objective is to implement the Gradient Boosting algorithm as a dependable housing decision-making aid for consumers. In our analysis and recommendations, we consider lifestyle, budgetary constraints, and the desire for a healthful living environment.

2. Methodology

Data Collection
Our analysis of the housing market in Bangalore is primarily based on a massive dataset containing over 13,000 observations from all over the city. Location, total square footage, number of bedrooms and restrooms, as well as other factors, were taken into account. This data was obtained from Kaggle, a well-known platform for disseminating a broad variety of datasets for the study of complex trends, such as those in real estate pricing[4].

Algorithm Used
Linear Regression, Random Forest Regressor, Decision Trees and Gradient Boost Regressor.

Algorithm Explanation
This study compares and contrasts four distinct methods for accurately assessing property values, each with its own set of advantages and distinguishing characteristics. The Linear Regressor, the Decision Tree Regressor, the Random Forest Regressor, and the Gradient Boost Regressor are examples of regressors. Each potential algorithm for completing this task possesses a unique set of benefits and capabilities.
1. Linear Regression: The objective of regression analysis is to determine the relationship between a singular predicted value (the dependent variable) and a number of numerical independent variables (the predictors).Due to Sir Francis Galton's groundbreaking genetics research, the term "regression" entered common usage in the late 19th century. Our contemporary understanding and application of statistical analysis is heavily dependent on these earlier works. Strangely, he discovered that extreme short or extreme tall fathers were more likely to produce offspring of average height. To describe this occurrence, he coined the term "regression to the mean"[13].
Scientists can use linear regression to describe complex systems by incorporating equations that

account for the influence of one variable on another. $y = \beta_0+\beta_1x_1+\beta_2x_2+...+\beta\square x\square$ to investigate the relationships between various variables. The impact of each predictor can be measured using beta coefficients (). By taking into account all of the contributing factors, these methods are especially useful for comprehending phenomena with multiple dimensions.

In this equation, y represents the anticipated property value, while x1, x2,..., xp are the input variables or characteristics (such as location or amenities). The coefficients 0 through p characterise the relationship between these input variables and the expected home price. These coefficients are estimated throughout model training.

2. Decision Tree: A decision tree is a model of a tree that has been trained on historical data to accurately predict future data. Decision Trees are constructed using the statistical approaches of Classification Trees and Regression Trees, which aim to optimise information acquisition via regression. The root node (sometimes referred to as the Parent Node) of a decision tree can be subdivided into offspring nodes, which can be further subdivided into Parent Nodes for subsequent Nodes. By establishing the nodes at the informative qualities as the maximisation of information acquisition, an objective function for enhancing the tree learning technique is formulated [4]. Modelling with a decision tree does not necessitate the use of unique equations or formulas.

3. Random Forest Regressor: (Breiman, 2001) Each tree in the Random Forest Regressor is constructed using a portion of the training data. To enhance the randomness of random forest construction, splitting operations on each node may evaluate either a random subset of available data or all features for optimal split (Ho, 1998). The extent of the arbitrarily selected subset is a user-specified hyperparameter. Frequently, the problem of overfitting reduces the effectiveness of individual decision trees. Random forest reaches a conclusion by combining these two forms of randomness. Methods such as random subset sampling and averaging can be utilised to reduce the effects of these errors and improve the precision of generalised estimation. In general, random forests are susceptible to developing bias, but variance is the most important aspect of bias to look out for [14].

4. Gradient Boost Regressor: In regression and classification tasks, the Gradient Boost Regressor (GBM) integrates the results of numerous decision trees to make precise predictions. Loss function, feeble learners, and additive model are the three pillars of GBM that contribute to its effectiveness.

Depending on the specifics of the circumstance, the Loss function may combine predefined metrics with human input to reach a conclusion.

In contrast to more complex processes, gradient boosting methods are sufficiently general to be readily combined with a wide diversity of loss functions. In the gradient boosting method depicted in Figure, poor pupils are represented by decision trees. When the finest portions of a tree are selected with ethics in mind, the tree flourishes. Although the intention is to secure the children's vulnerability, this strategy may be inherently exploitative[15].

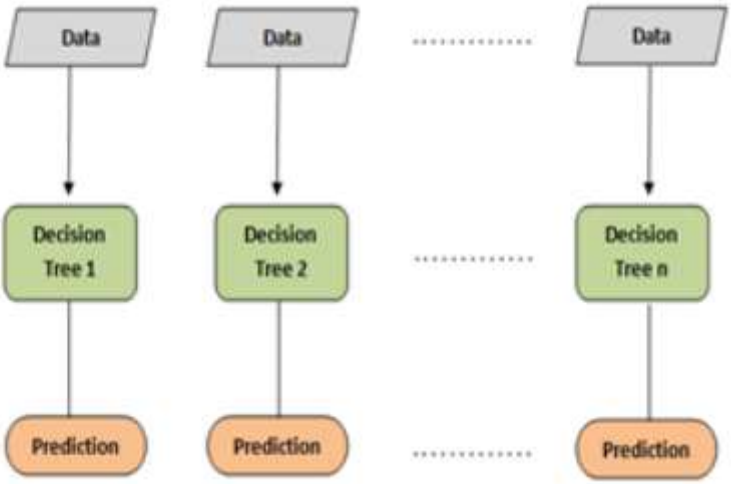
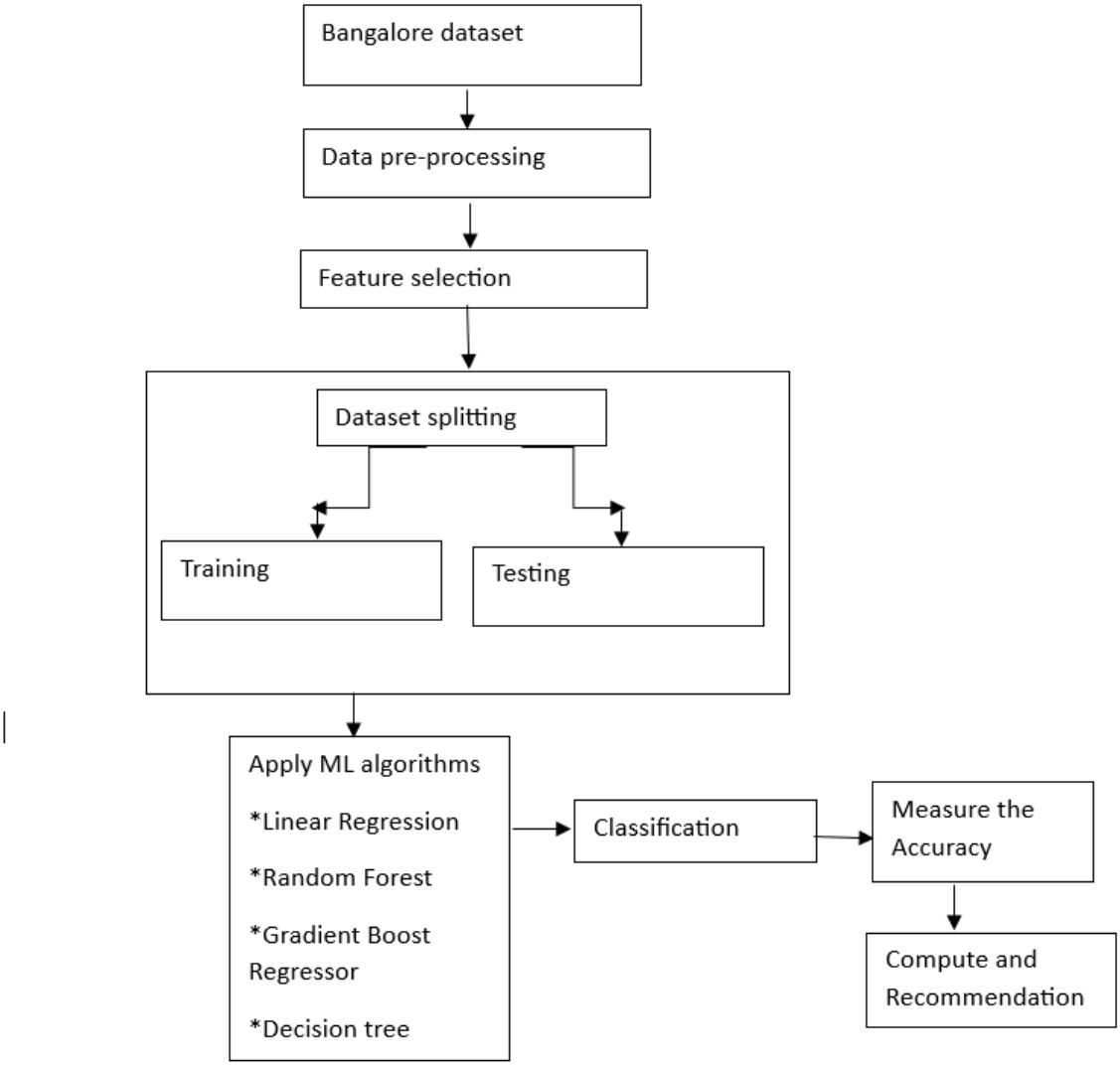


Fig 1: Gradient Boost Algorithm

SYSTEM ARCHITECTURE DIAGRAM



Pseudocode

Here are the process steps for the pseudo code provided:

Step 1: Import header files as needed for implementation for further work.

Step 2: Read the dataset from the shared folder.

step 3: Remove any unwanted columns from the dataset that are not relevant for house price computation and recommendation.

step 4: Handle missing values in the dataset by dropping rows with null values.

step 5: To determine the price per square foot a basic calculation is required. Simply divide the price column by the square footage column and voila! It is feasible to append a fresh column into the dataset that encompasses this particular set of details.

step 6: The detection and removal of outliers within the dataset is paramount for maintaining high performance levels of the model.

step 7: Save the cleaned dataset to a new file for further processing and analysis.

step 8: Split the cleaned dataset into features (X) and target variable (y), and further split them into training and testing sets using train_test_split function.

step 9: Train multiple machine learning algorithms (linear regression, random forest, decision tree, gradient boosting) using the training set, make predictions on the testing set, and calculate the r2 score for each algorithm's predictions.

step 10: Identify the algorithm with more accuracy as the best-performing algorithm for house price computation and recommendation.

step 11: Perform label encoding on categorical variables, such as converting the location column from text to numeric representation for further processing.
step 12: Recommend Houses based on Predicted Values: Use the selected algorithm to compute the price of a new instance (house) by inputting its features (bedrooms, bathrooms, sqft, encoded location). Display the computed price and recommendation.

3. Result And Discussion

In this study, we place no emphasis on the application of the Gradient Boosting Regressor model to real estate price forecasting and recommendation. Comparison to other methods such as Linear Regression, Random Forest, and Decision Tree validates the veracity of the model.
With 99.94% precision, the Gradient Boosting Regressor was by far the most accurate model. The Gradient Boosting Regressor was chosen as the primary model due to its increased precision.
Figure1 is a visual representation of the comparison between the performance of the Gradient Boosting Regressor model and other machine learning techniques discussed in [3].

	Model	r2_score
3	Decision Tree	0.992592
0	Random Forest	0.998363
1	Gradient Boost Regressor	0.999472
2	Linear Regression	1.000000

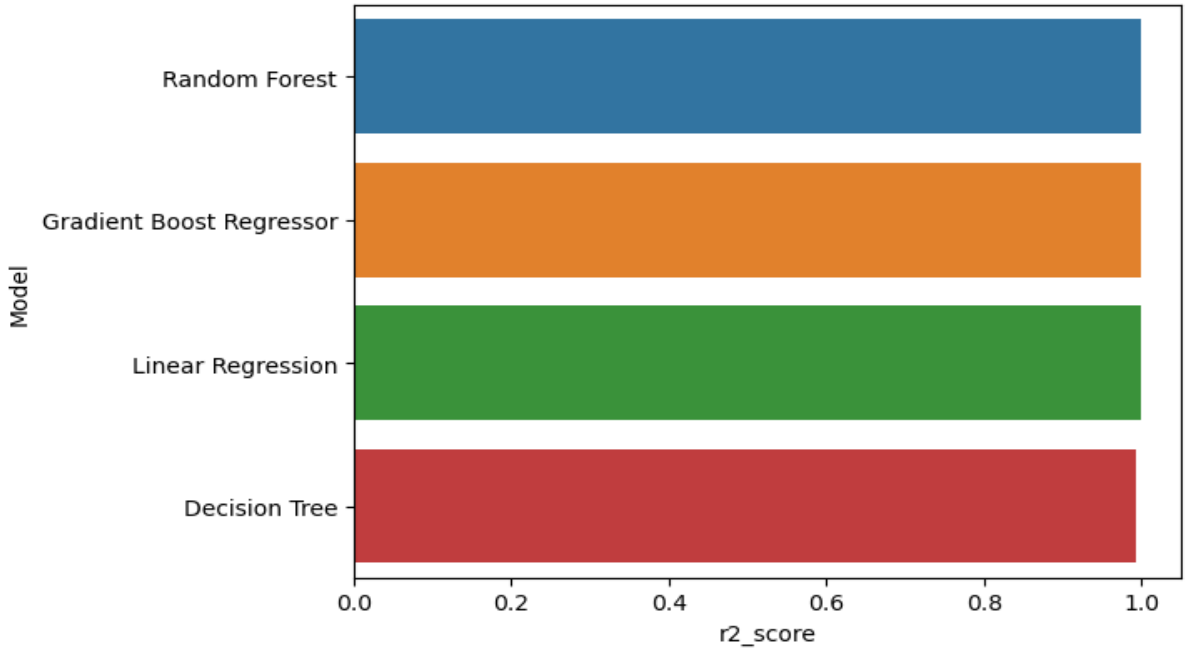
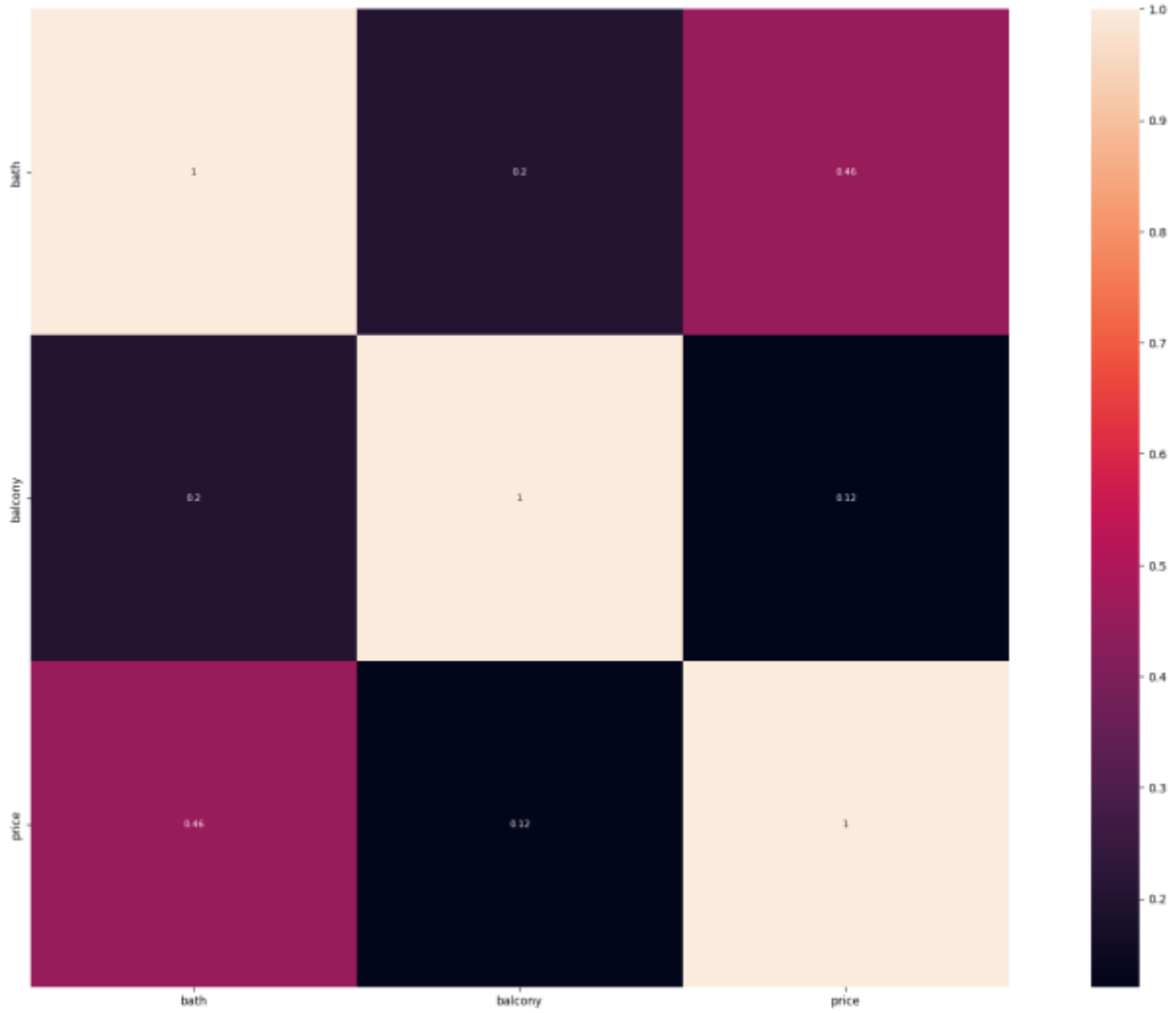


Fig 2: Comparitive Analysis

Using a correlation matrix is one of the final steps in our data set analysis. A correlation matrix provides a table containing the coefficients of correlations between random and other values in order to gain insight into the strength of the link between two or more variables. We created a seaborne heatmap to visually represent the complex statistical interplay at work in this investigation, drawing attention to critical areas within the vast amounts of data by judiciously applying colour shading. Figure 2 depicts how Seaborn heatmaps are lauded by data science experts for their aesthetic appeal and ability to convey fundamental interpretations of complex analytics results via visualisations alone.

Fig 3: Heatmap/correlation matrix



We Found Significant Correlations Between Bhk, Location, And Bath. It Was Determined That Additional Characteristics With Scores Greater Than 0.5 Had Sufficient Correlation To Be Used As Predictors In The Cost Analysis. Consequently, We Will Use This Method To Predict Future Pricing.

In This Instance, A Website Will Function As The GUI Interface For The Project. Location, Number Of Bedrooms, Square Footage, And Number Of Restrooms Are Required Inputs For Calculating Prices And Recommending Features On The Website[10].

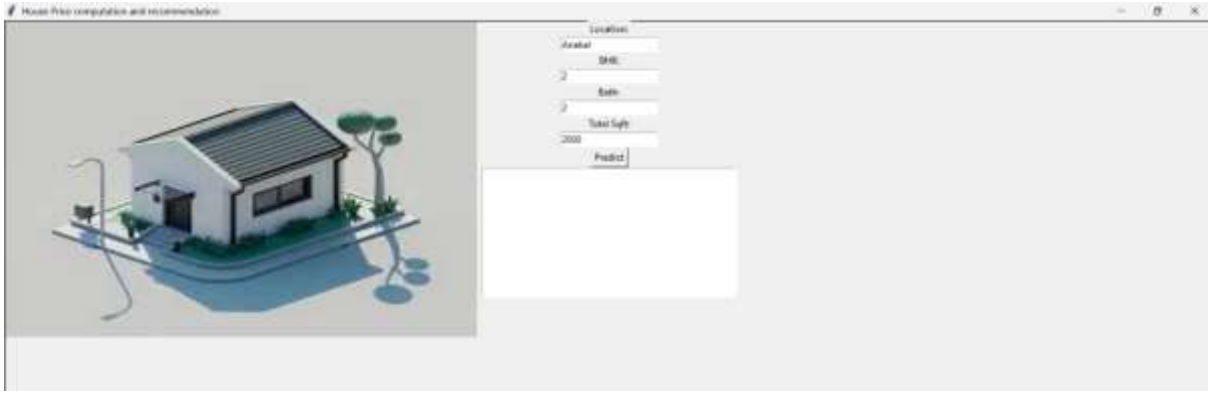


Fig 4 : GUI Frontend

Our GUI system design framework includes a "Predict" command icon that, when activated, provides users with instantaneous price range estimates for properties that meet the criteria they

specified at the outset of their search for cost-effective solutions.

Fig 5: Result After Prediction

4. Conclusion

This undertaking has taught us a great deal about machine learning model construction, regression, and power transformer models. Our research provided insight into their evolution and potential application within the context of property valuation and advice.

Several methods, including linear regression, random forest regressor, decision tree, and gradient boost regressor, have been extensively studied in an effort to improve the accuracy with which future housing prices can be predicted. These algorithms were developed and evaluated to determine how accurately they could predict and recommend housing market price fluctuations.

Our experiment demonstrates that Gradient Boost Regressor provides the highest level of accuracy, with a 99.94% success rate, after extensive research and comparison. Our research compiles graphical representations of comprehensive analyses of the efficacy of each algorithm. The comparison highlights the advantages and disadvantages of each option.

Our primary focus was on real estate price forecasts, but we also provided valuation estimates and professional guidance. Given the relevant variables, we used the Gradient Boost regressor model to accurately estimate house values. Those looking for a new home may also rely on our trustworthy recommendations. As a consequence, individuals will be better equipped to make informed decisions. Our results are significant, especially the use of gradient regression with high precision and the graphical representation of technique performance. This research was conducted to collect as much data as feasible for future use in regression models and power transformers regarding the computation and recommendation of property prices.

5. References

1. Anand g. Rawool, dattatray v. Rogye, sainath g. Rane, dr. Vinayk a. Bharadi, "House Price Prediction Using Machine Learning.", IRE Journals , Volume 4 ,Issue 11,2021
2. Durganjali, P., and M. Vani Pujitha. "House resale price prediction using classification algorithms." In 2019 International Conference on Smart Structures and Systems (ICSSS), pp. 1-4. IEEE, 2019.

3. Kaushal, Anirudh, and Achyut Shankar. "House Price Prediction Using Multiple Linear Regression." In Proceedings of the International Conference on Innovative Computing & Communication (ICICC). 2021.
4. Amey Thakur, Mega Satish. "BANGALORE HOUSE PRICE PREDICTION." , International Research Journal of Engineering and Technology (IRJET), Volume 08, Issue 09, 2021.
5. Mysore, Sumanth, Abhinay Muthineni, Vaishnavi Nandikandi, and Sudersan Behera. "Prediction of house prices using machine learning." Int. J. Res. Appl. Sci. Eng. Technol 10, no. 6 (2022): 1780-1785.
6. Sifei Lu, Zengxiang Li, Zheng Qin , Xulei Yang , Rick Siow Mong Goh - "A hybrid regression technique for house prices prediction", IEEE, 2017.
7. Madhuri, CH Raga, G. Anuradha, and M. Vani Pujitha. "House price prediction using regression techniques: A comparative study." In 2019 International conference on smart structures and systems (ICSSS), pp. 1-5. IEEE, 2019.
8. Viktorovich, Parasich Andrey, Parasich Viktor Aleksandrovich, Kaftannikov Igor Leopoldovich, and Parasich Irina Vasilevna. "Predicting sales prices of the houses using regression methods of machine learning." In 2018 3rd Russian-Pacific conference on computer technology and applications (RPC), pp. 1-5. IEEE, 2018.
9. Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita. "House Price Prediction using various Regression Analysis and Particle Swarm Optimization".(IJACSA) International Journal of Advanced Computer Science and Applica- tions, Vol. 8, No. 10, 2017.
10. Gang-Zhi Fan, Seow Eng Ong and Hian Chye Koh, "Determinants of House Price: A Decision Tree Approach", Urban Studies, Vol. 43, No. 12, November 2006, PP.NO.2301- 2315.
11. Ong, S. E., Ho, K. H. D. and Lim, C. H., "A constant quality price index for resale public housing flats in Singapore", Urban Studies, 40(13), 2003, pp. 2705 –2729.
12. Berry, J., McGreal, S., Stevenson, S., "Estimation of apartment submarkets in Dublin, Ireland", Journal of Real Estate Research, 25(2), 2003, pp. 159–170.
13. Chouthai, Atharva, Mohammed Athar Rangila, Sanved Amate, Prayag Adhikari, and Vijay Kukre. "House Price Prediction Using Machine Learning." International Research Journal of Engineering and Technology (IRJET) 6, no. 03 (2019).
14. Jha, Shashi Bhushan, Radu F. Babiceanu, Vijay Pandey, and Rajesh Kumar Jha. "Housing market prediction problem using different machine learning algorithms: A case study." arXiv preprint arXiv:2006.10092 (2020).
15. Monika, R. "House Price Forecasting Using Machine Learning Methods." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12, no. 11 (2021): 3624-3632.