Asian Journal of Applied Science and Technology (AJAST)
Volume 7, Issue 1, Pages 41-54, January-March 2023

Machine Learning Based House Price Prediction Using Modified Extreme Boosting

N. Ragapriya[1*], T. Ananth Kumar[2], R. Parthiban[3], P. Divya[4], S. Jayalakshmi[5] & D. Raghu Raman[6]

[1]UG Student, [2,3,6]Associate Professor, [4,5]Assistant Professor,
[1-6]Department of Computer Science & Engineering, IFET College of Engineering, Villupuram, India.
*Corresponding Author (N.Ragapriya) - Email: ragapriyavcn14@gmail.com**

DOI: https://doi.org/10.38177/ajast.2023.7105

Article Received: 14 January 2023 *Article Accepted: 27 February 2023* *Article Published: 07 March 2023*

ABSTRACT

In recent years, machine learning has become increasingly important in everyday voice commands and predictions. Instead, it provides a safer auto system and better customer assistance. As a result of all that has been demonstrated, ML is a technology that is becoming more and more popular in a range of industries. To gauge changes in house values, the House Price Index is frequently employed (HPI). Due to the substantial correlation that exists between property prices and other variables, such as location, region, and population, the HPI on its own is not sufficient to accurately forecast a person's house price. Some studies have successfully predicted house prices using conventional machine learning techniques, but they seldom evaluate the efficacy of different models and ignore the more complicated but less well-known models. We proposed Modified Extreme Gradient Boosting as our model in this study due to its adaptive and probabilistic model selection process. Feature engineering, hyperparameter training and optimization, model interpretation, and model selection and evaluation are all steps in the process. Home price indices, which are frequently used to support real estate policy initiatives and estimate housing costs. In this project, models for forecasting changes in home prices are developed using machine learning methods.

Keywords: Home price; Location; Square footage; Modified extreme gradient boosting.

1. Introduction

The housing market is a subset of the real estate sector, which is essential to the overall health of any economy. Because the ownership of a home is seen as a status symbol in many parts of the world, a number of newly employed people have set this as a professional objective for themselves. Despite this, investors are drawn to the real estate market because they see it not as a commodity but rather as an opportunity [1].

A thriving real estate market, which must necessarily include a robust housing market, is necessary for an economy that is expanding. Because of the status symbol that it represents, becoming a homeowner is a goal for many young professionals in many countries. This is because home ownership represents status. Investors, on the other hand, are drawn to the housing market because they do not view property as a commodity but rather [2] as a source of potential profit. Homebuyers and investors alike typically enter the real estate market with the expectation of making a profit from subsequent price appreciation. There is an inverse relationship between the price of homes and the percentage of people who own their homes. In the past, the majority of research has focused on nations that have high rates of homeownership, particularly those nations that are still on the rise economically [3].

It is essential for people to have access to housing that is not only affordable but also meets their basic needs because the cost of housing has a significant influence on the long-term viability of the market. The degree to which housing can be purchased on a reasonable budget is a critical factor in determining whether or not it is a good long-term investment strategy. The stock market, interest rates, and the currency exchange market are all significantly more volatile than the real estate market, which is significantly less volatile. The fluctuations in home prices have a significant impact on the real estate market, which in recent years has emerged as [4] one of the most fruitful fields for investment, particularly over the course of the past 15 years. In recent years, one of the most widely discussed topics related to real estate has been the methodology behind the pricing of properties. As a direct result of this matter, a variety of stakeholders, including residential investors, real estate investment trusts, individual investors, and officials from various government [5] agencies, have been prompted to make projections regarding the future path of home values. These individuals have employed a large number of different methods in order to achieve their objective. Since the beginning of the Industrial Revolution, urban populations have increased at an exponential rate, [6] which has resulted in a severe lack of available housing. This is due to the rapid urbanisation that is taking place in every region of the world at the present time. As more time has passed, a variety of facets of this problem have become more apparent. This book investigates the widespread housing problems that plagued Germany's largest cities in the 1980s, when the nation was still in the process of developing its infrastructure. During this time period, society as a whole started to come to terms with the housing shortage that was a direct result of economic inequality, as well as the subsequent requirement for more [7] social housing. These traits have also been identified as problems in developing countries with a more recent urban population, where cities are still in their early stages of development.

For example, the issue of housing that is not up to standard in India has been the subject of a number of studies. While Abhay was presenting data on the number of homes built in Karnataka over the course of the previous decade, he mentioned that despite the housing shortage problem encouraging the building supply, low-quality housing is widespread throughout the city (2001-2011). When Abhay was talking about the number of residential construction projects that took place in Delhi over the course of the examined decade, he brought this up [8]. Many people point to the disproportionately high number of old houses as a major factor that contributes to the shortage of housing options. It is challenging to estimate the value of a home because of the intricate connections that exist between a property's physical characteristics, the neighbourhood, and its location.

In particular, the home price forecast model has received a significantly smaller amount of attention in the existing body of literature in comparison to more conventional approaches to the solution of this problem. In spite of the fact that there is widespread consensus that this constitutes an urgent problem that needs to be fixed, this keeps occurring. Machine learning has emerged as an important prediction method as a result of the rise of Big Data [9]. This has made it possible to make more accurate property price forecasts based solely on features of the property, rather than on historical data. Although a number of studies have been conducted to investigate this question and found that machine learning is effective, the vast majority of these studies have merely compared the performances of different models without taking into consideration how to combine different machine learning models in the most effective way [10]. A modification of the algorithm known as extreme gradient boosting is used in this research to perform price forecasting for residential real estate. As a result, becoming more familiar with regression methods within machine learning is intended to be the end result of this project. In order to achieve optimal performance, it is necessary to process the datasets that have been provided. Because the value of a house is determined by the characteristics that are unique to it, [11] we need to first establish which of those characteristics are essential before we can eliminate the ones that aren't important and arrive at an accurate estimate. The data are skewed as a result of the fact that not all homes would have access to these upgrades; consequently, the price of houses would not change in the same way without them.

The main objective of this project is to predict the house price using Modified Extreme Gradient Boosting algorithm. It is used the predict the price using the area type, location, BHK etc., The Modified XGBoost is measured for accuracy and performance against several algorithms.

2. Literature Survey

The most precise machine learning (ML) models for predicting property values were studied by Park et al. [4]. For this, they examined a sample of 5359 Virginian row homes. They employed two methods: the first, the RF approach, was utilised to address a classification issue, while the second, the nave Bayesian algorithm, addressed a regression issue. According to the findings, the RIPPER algorithm significantly enhanced price prediction. Several different approaches to machine learning were investigated by Banerjee et al. [12] for the purpose of predicting whether future real estate prices will go up or down. The RF method was found to have the greatest amount of overfitting but was also the method with the highest degree of accuracy. However, the SVM technique was the most trustworthy because it remained unchanged over the course of the study. Kok et al. [13] investigated the usefulness of several different machine learning strategies in the context of real estate appraisals. These methods included the RF approach, and the GBR method. The findings demonstrated that the XGBM algorithm achieved the highest levels of performance overall. Ceh et al. [14] compared the hedonic price model to the RF algorithm in order to determine which technique would result in more accurate price forecasts. The authors surveyed a total of 7407 different homes in the Ljubljana area between the years of 2008 and 2013. As (Slovenia). a result of the outcomes, it was clear that the RF model had a superior performance in terms of prediction 5. Using supervised learning methods, Hu et al. [15] investigated how accurate it is to predict how much it will cost to rent a house in Shenzhen (China). The authors used the multilayer perceptron neural network (MLP-NN), the k-NN, the RF, the ETR, the GBR, and the SVR algorithms in their research. The findings indicated that both the RF and ETR algorithms operated in a more predictable manner.

S. Lu et al. [16] created a model with improved results for evaluating London's spatial closeness and better estimate performance for property values. These two geographic traits are two examples of non-linear phenomena that defy linear quantification. Elham Alzain et al. [17].'s goal is to use an ANN-based prediction model to forecast Saudi Arabia's future home values. In Riyadh, Jeddah, Dammam, and Al-Khobar, four significant Saudi Arabian cities, the dataset was gathered from Aqar. The findings indicated that the experimental and projected values had a good degree of agreement. The employment of ANN in this results in an accuracy of 80%. Classification algorithms were employed by P. Durganjali et al. [18] to predict the cost of resale properties. The selling price of a property is predicted in this study using a variety of classification methods, including Leaner regression, Decision Tree, K-Means, and Random Forest. A home's price is influenced by its physical attributes, its geographic location, and even the state of the economy. Here, we use these techniques, utilise RMSE as the performance matrix for different datasets, and find the most accurate model that properly predicts improved results.

A hybrid regression technique was created by Sifei Lu et al. [19] to forecast housing prices. A tiny dataset and data characteristics are used in this study to test the creative feature engineering approach. Recent Kaggle Competition "House Price: Advanced Regression 6 Methods" entries adopted the suggested methodology as their fundamental building block. In view of the readers' financial limitations, the article's goal is to estimate reasonable pricing for the customers. Fletcher et al. [20] (2000) investigate whether it is preferable to employ aggregate or disaggregated data when hedonic analysis is used to forecast housing price. It is discovered that several features' hedonic pricing coefficients vary significantly depending on age, location, and kind of property. However, it is believed that using an aggregate method, this may be properly simulated. The individual external impacts of factors like environmental attributes on housing prices have also been estimated using the hedonic price model. For instance, several studies have used the hedonic price model to measure how much noise and air pollution affect home prices.

2.1. Problem Statement

The current technique can forecast some straightforward and modest homes. It also forecasts the accuracy of different algorithms. It predicts price using a single parameter. The system as it is now does not benefit from its current training phase. They are unable to forecast a mansion of luxury. It is unable to suggest the optimal algorithms [21]. Several parameters cannot be used for prediction. The loss function with regularisation is the objective function that has to be minimised at iteration t:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\Big(\overset{\text{Real value (label) known from the training data-set}}{\overset{\Uparrow}{y_i}}, \underbrace{\hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)}_{\Downarrow \atop \text{Can be seen as } f(x + \Delta x) \text{ where } x = \hat{y}_i^{(t-1)}}\Big) + \Omega(f_t)$$

The XGBoost goal is obviously a function of functions and "cannot be maximized using typical optimization methods in Euclidean space," as stated.

From, we get the fundamental linear approximation of the function f(x simplest) as follows:

$$f(x) \approx f(a) + f'(a)\underbrace{\boxed{(x-a)}}_{\Delta x = f_t(\mathbf{x}_i)}$$

To represent the initial function, only a function of x may be used:

The initial function can only be expressed as a function of x.

In order to convert a function f(x) into the simplest function of x that can be found near a certain point, use Taylor's theorem. Prior to the application of the Taylor approximation, the objective function f(x), in which x represented the sum of the t CART trees, was a function of the currently-selected tree (step t). In this particular scenario, the anticipated outcome from step t-1 is an, the loss function is denoted by the expression f(x), and the variable x denotes the additional student who needs to be added in step t.

Given the information that was presented earlier, it is possible to optimise in Euclidean space by defining the objective (loss) function as a straightforward function of the new learner that was introduced at each iteration [22]. This makes it possible to perform optimization in Euclidean space (t). As was mentioned earlier, the additional learner that needs to be included in step (t) in order to eagerly reduce the target is a stand for the prediction that was made in steps (t-1) and (x-a). This step is required in order to eagerly reduce the target. When the Taylor approximation of the second order is utilised:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2} f''(a)(x-a)^2$$

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

XGBoost objective using second-order Taylor approximation-

Where:

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \text{ and } h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$$

The loss function's first and second order gradient statistics-

At least, if we eliminate the constant components, we are left with the simplified objective to minimise at step t as follows:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

XGBoost simplified objective-

Our next objective is to identify a learner that minimises the loss function at iteration t since the aforementioned is a sum of straightforward quadratic functions of a single variable that can be minimised using well-known methods.

$$argmin_x \; Gx + \tfrac{1}{2}Hx^2 = -\tfrac{G}{H}, \;\; H > 0 \qquad \min_x \; Gx + \tfrac{1}{2}Hx^2 = -\frac{1}{2}\frac{G^2}{H}$$

Minimizing a simple quadratic function-

Notice how the following scoring function resembles the "basic quadratic function solution" above when using the authors' method to "assess the quality of a tree structure q":

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\overset{\text{instances mapped to leaf}_j}{\boxed{(\sum_{i \in I_j} g_i)^2}}}{\sum_{i \in I_j} h_i + \lambda} + \gamma T.$$

The tree learner structure q scoring function

$$y \ln(p) + (1-y)\ln(1-p) \text{ where } p = \frac{1}{(1+e^{-x})}$$

Binary classification using the Cross Entropy loss function, where p stands for the probability score and y for the actual label, both of which are in the range of 0 to 1 [23]. The total of the CART tree learners makes up the model's output x. We must ascertain the gradient's and hessian's first and second derivatives with respect to x in order to minimise the log loss objective function. For example, this post on Stats Stack Exchange may teach you that the formula for the gradient is (p-y) and the formula for the hessian is (p-*) (1-p).

3. Proposed System

This work is used to predict the house price in safe and secured manner. It can predict the house price for great geo location. It uses Extreme Gradient Boosting for better accuracy. This algorithm uses linear function, likelihood etc., It helps the prediction more accurate and easier.

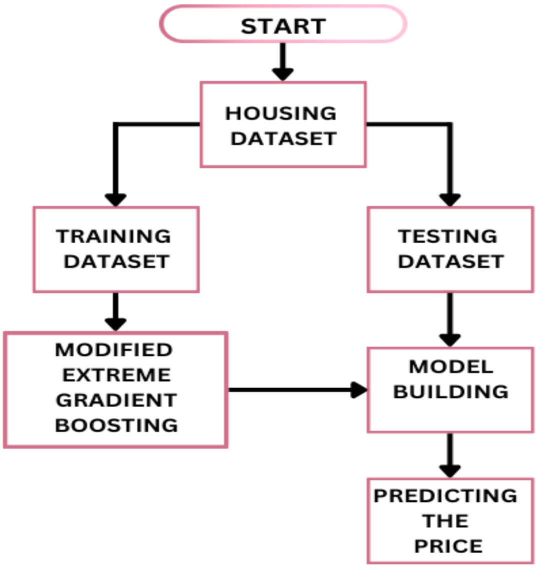


Figure 1. Flow Diagram

The data is collected from the database and the data is pre-processed such noise reduction, anomaly detection, missing value finding etc., The pre-processed data is trained using various algorithms and the model is tested using testing data. Then build the model using modified extreme boosting algorithm. Connect the model with flask and execute. Get the input from the using in the website and extract the keywords from the input. Then predict the value for the data given.

3.1. System Architecture

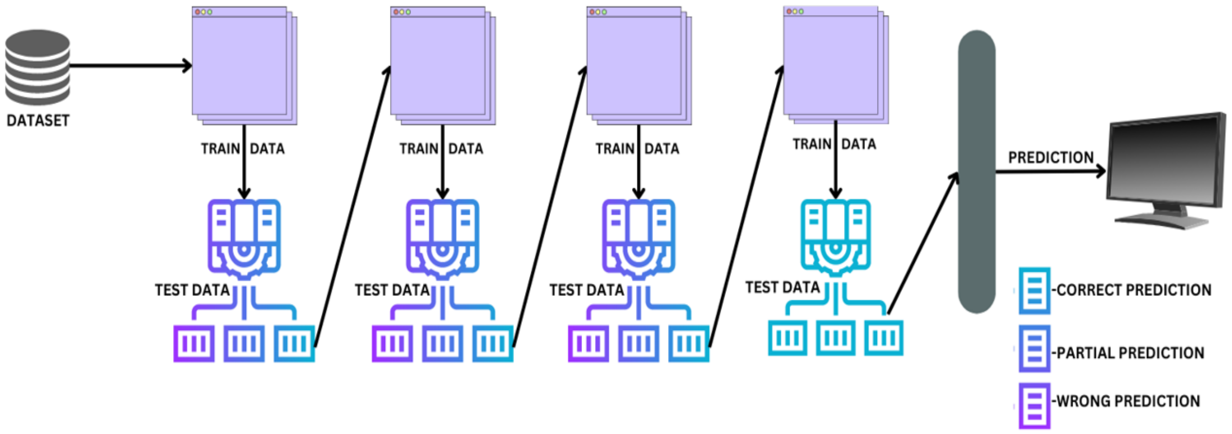


Figure 2. Proposed architecture

Algorithm 1. Modified Extreme Boosting

```
    Initialization:
    STEP 1:Given training information from the instance space S = (x1, y1),..., (m, ym) where xi EX and y; Y = -1,+1.
    STEP 2: start the distribution off. $D_1$(i) =.
    Using the algorithm for t = 1,..., T, do
    Utilizing the distribution Dt, train a weak learner ht: X R.
    Zt is a normalisation factor chosen to create the distribution of Dt+1, and its formula is
    Dt+1(i) = Dt(i)e-tyiht(xi) Zt.
    f(x) = oatht(x) and H(x) = sign(f(x)) are the final scores.
```

3.2. Modified Extreme Boosting Algorithm

A. General Parameters

silent: Zero is the default value. For printing running messages, you must specify 0; for silent mode, you must specify 1.

booster: GBTree is the default selection. The booster to employ must be specified: Tree-based GBTree or linear GBLinear (linear function).

buffer: It is automatically set by the XGBoost algorithm; user input is not required.

num feature: XGBoost Algorithm sets this value automatically; human input is not required.

B. Booster Parameters

ETA: 0.3 is the default setting. To avoid overfitting, you must specify the step size shrinkage utilised in an update. You can immediately obtain the weights of new features after each boosting stage. In order to increase the conservatism of the boosting process, eta actually reduces the feature weights. It ranges from 0 to 1. The model is more resistant to overfitting if the eta value is low.

GAMMA: A default value of 0 has been set. On a leaf node of the tree, you must indicate the minimal loss reduction necessary to create another division. The algorithm will be more conservative the larger it is. From 0 to is the range. The algorithm is more conservative the larger the gamma.

MAX DEPTH: A value of six is used by default. A tree's maximum depth must be specified. 1 to make up the range.

MIN CHILD WEIGHT: By default, the value is set to 1. The bare minimum instance weight (hessian) required in a child must be specified. if a leaf node emerges from the tree partitioning step. then with instance weight added together less than the minimum kid weight. The construction process will then give up on more partitions.

corresponds to the bare minimum of instances required to be in each node in linear regression mode. The algorithm will be more conservative the larger it is. From 0 to is the range.

MAX DELTA STEP: 0 is the default setting. The largest delta step that can be used to estimate each tree's weight. There is no constraint if the value is set to 0. It can assist make the update step more cautious if it's set to something positive. Although it is not typically required, this parameter may be useful in logistic regression. especially when there is a severe imbalance in the class. Setting it to a number between 1 and 10 could aid in updating control. From 0 to is the range.

SAMPLE: 1 is set as default option. The subsample ratio for the training instance must be given. Half of the data instances will have been randomly chosen by XGBoost if you set it to 0.5. Overfitting must be prevented by allowing trees to develop. It is between 0 and 1.

COLSAMPLE BYTREE: The value is set to 1 by default. When building each tree, you must give the subsample ratio of columns. It ranges from 0 to 1.

C. Linear Booster Specific Parameters

These XGBoost Algorithm Linear Booster Specific Parameters are used.

LAMBDA AND ALPHA: These are terms for weight regularisation. Alpha is supposed to be 0, while lambda's default value is 1.

LAMBDA BIAS: This term has a default value of 0 and is an L2 regularisation term on bias.

D. Learning Task Parameters

The XGBoost algorithm's learning task parameters are listed below.

BASE SCORE: 0.5 is the default value for this field. The initial prediction score and global bias for each occurrence must be specified.

OBJECTIVE: Reg: linear is the default value set. You must be clear about the kind of learner you require. This comprises Poisson and linear regression, among others.

EVAL METRIC: The evaluation metrics for the validation data must be specified. Moreover, a default metric will be chosen in accordance with the goal.

SEED: In order to reproduce the same set of outputs, you must specify the seed here as usual.

3.3. Model Building

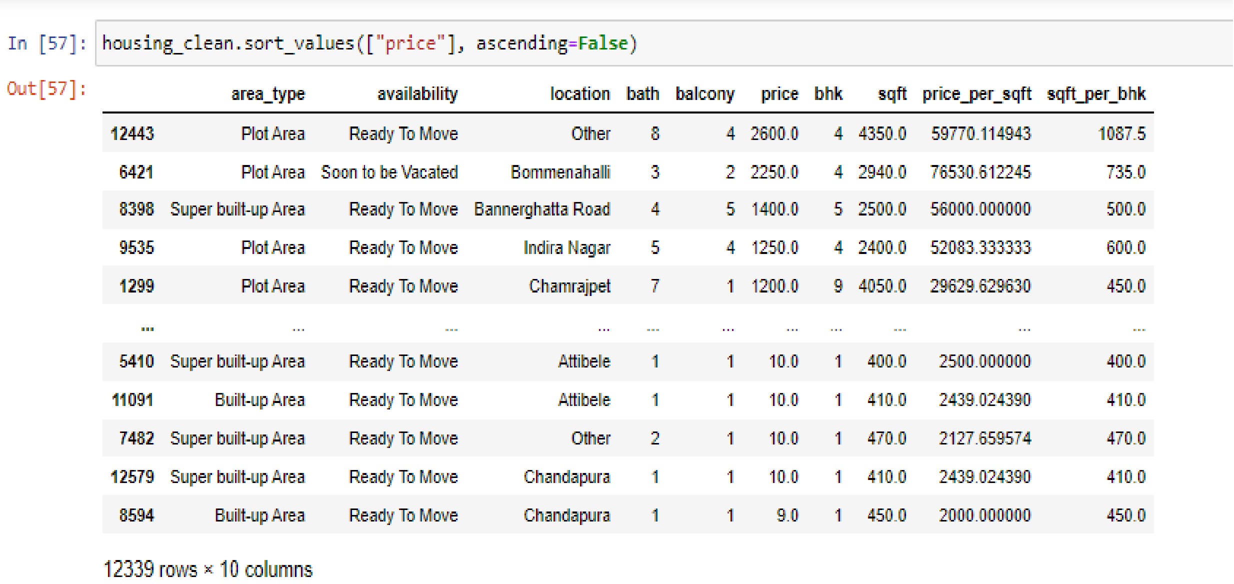
Two datasets are required when building a machine learning model: one for training and one for testing. However, there is now only one. Let's divide this in two according to an 80:20 ratio. Create two columns for features and labels in the data frame. Here, the train-test split function from Sklearn was loaded. After that, divide the dataset using it. Moreover, the dataset is split into two halves with test size = 0.2: 20% for the test and 80% for the train. Using a random number generator that is seeded by the random state parameter, the dataset is partitioned. The method returns four datasets. These were designated as test x, test y, train x, and train y. To match the data to several decision trees, use Modified Extreme Boosting. Finally, tell the fit method to pass trains x and y to train the model. After training, the model must be tested using test data. One of the most potent techniques used in ML for regression issues is modified extreme boosting. The supervised algorithm class includes the random forest. This procedure is run in three stages: the first involves assigning the independent variable weight, the second involves creating the forest from the provided dataset, and the third involves making predictions from the regressor.

Table 1. Algorithm used and its function

ID	Model	Function
1	Linear Regression	sklearn.linear_model.LinearRegression
2	Random Forest Regressor	sklearn.ensemble.RandomForestRegressor
3	Gradient Boosting Regressor	sklearn.ensemble.GradientBoostingRegressor
4	Ridge Regression	sklearn.linear_model.Ridge
5	Lasso Regression	sklearn.linear_model.Lasso
6	Ada Boosting Regression	sklearn.ensemble.AdaBoostRegressor
7	Decision Tree Regression	sklearn.tree.DecisionTreeRegressor
8	Modified Extreme Boosting	XGBRegressor()

4. Result and Discussion

The figure 3 shows the data and parameter present in the dataset. It also shows the columns and the rows in the csv file.



```
In [57]: housing_clean.sort_values(["price"], ascending=False)
```

```
Out[57]:
                area_type        availability          location bath balcony  price bhk    sqft price_per_sqft sqft_per_bhk
12443           Plot Area       Ready To Move             Other    8       4 2600.0   4  4350.0   59770.114943       1087.5
 6421           Plot Area  Soon to be Vacated      Bommenahalli    3       2 2250.0   4  2940.0   76530.612245        735.0
 8398  Super built-up Area      Ready To Move  Bannerghatta Road    4       5 1400.0   5  2500.0   56000.000000        500.0
 9535           Plot Area       Ready To Move       Indira Nagar    5       4 1250.0   4  2400.0   52083.333333        600.0
 1299           Plot Area       Ready To Move         Chamrajpet    7       1 1200.0   9  4050.0   29629.629630        450.0
  ...                 ...                 ...               ...  ...     ...    ... ...     ...            ...          ...
 5410  Super built-up Area      Ready To Move           Attibele    1       1   10.0   1   400.0    2500.000000        400.0
11091       Built-up Area       Ready To Move           Attibele    1       1   10.0   1   410.0    2439.024390        410.0
 7482  Super built-up Area      Ready To Move              Other    2       1   10.0   1   470.0    2127.659574        470.0
12579  Super built-up Area      Ready To Move         Chandapura    1       1   10.0   1   410.0    2439.024390        410.0
 8594       Built-up Area       Ready To Move         Chandapura    1       1    9.0   1   450.0    2000.000000        450.0

12339 rows × 10 columns
```

Figure 3. Dataset Description

The figure 4 shows the graph about the data present in the dataset. The ratio of the data is represented as histograms and bar graph. The bar graph represents the area_type present in the dataset and its count. The first histogram represents the sqft_per_bhk and its count and density (The weightage given for that parameter). The second histogram represents the price_per_sqft. The third histogram represents the sqft present in the dataset. The fourth histogram represents the total_sqft of the house is the buy. The last histogram represents the price in the dataset.

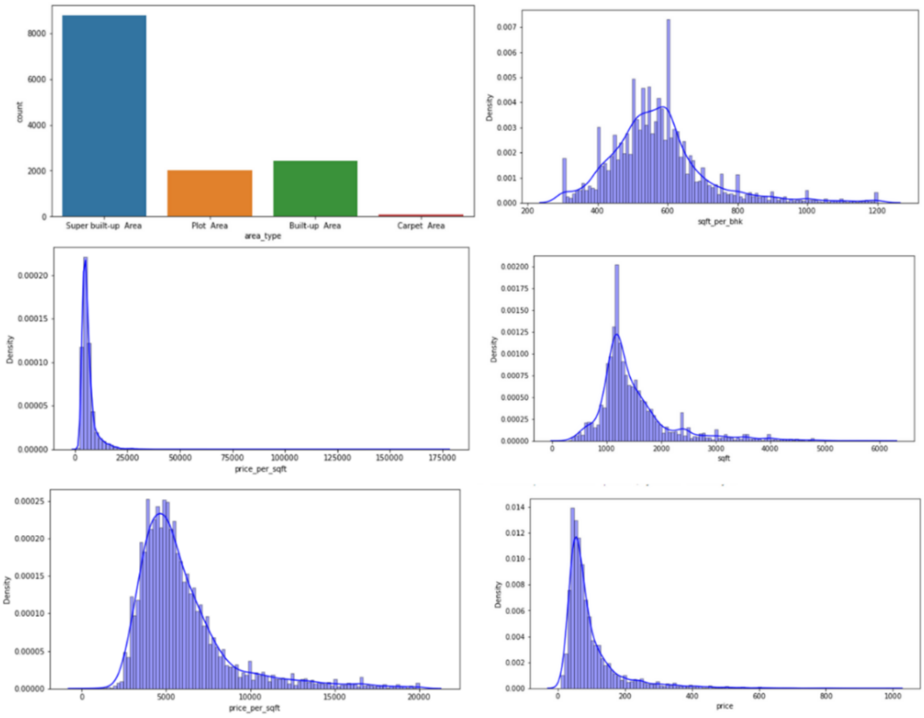
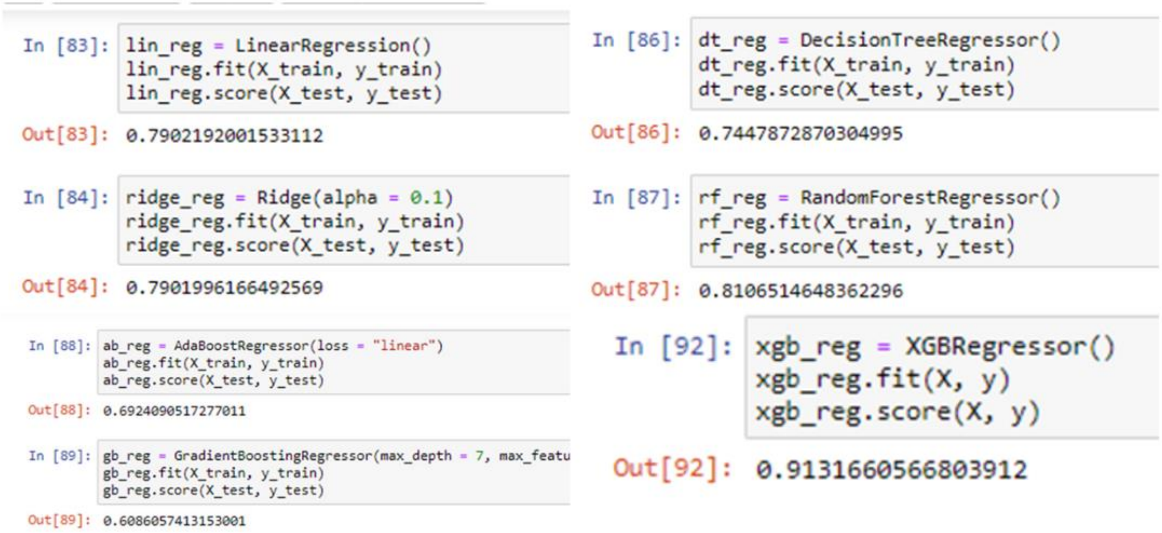


Figure 4. Data Graph

The figure 5 shows the various algorithm used and its accuracy. The algorithm used for Linear Regression, Ridge Regression, AdaBoost Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression and Modified Extreme Boosting Algorithms are used to compare the accuracy with Modified XGBoost.



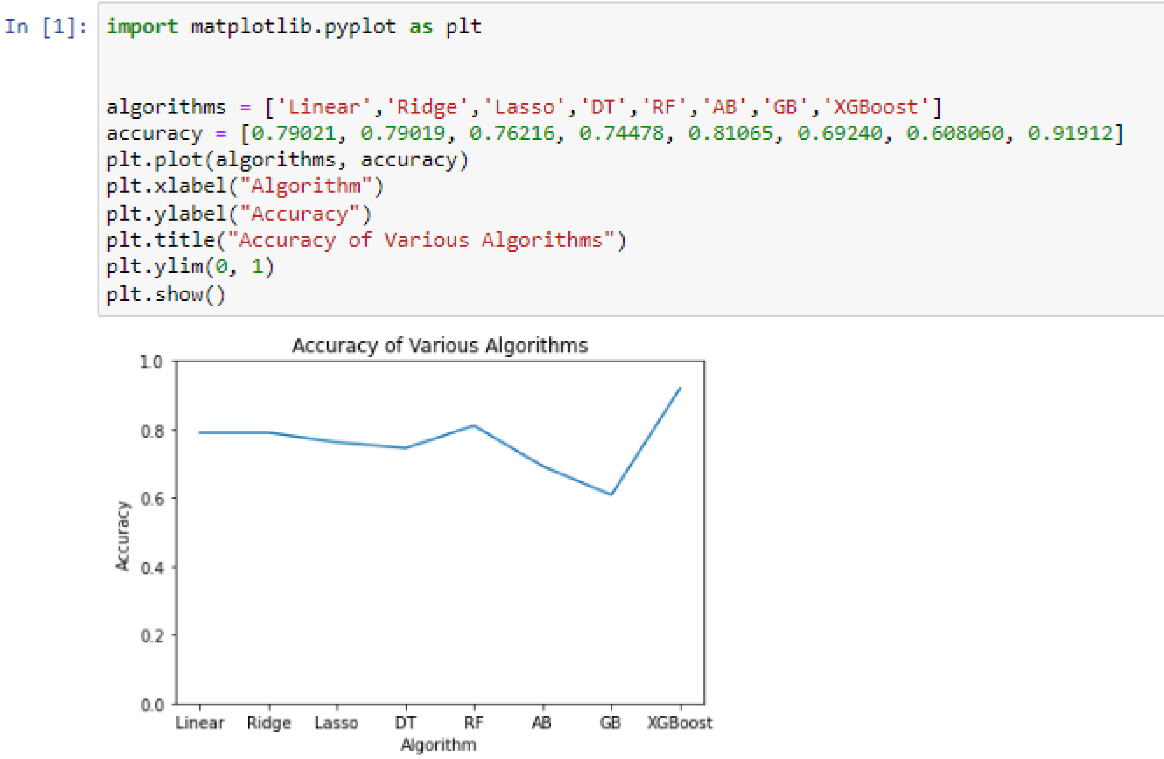
```
In [83]: lin_reg = LinearRegression()
         lin_reg.fit(X_train, y_train)
         lin_reg.score(X_test, y_test)

Out[83]: 0.7902192001533112

In [84]: ridge_reg = Ridge(alpha = 0.1)
         ridge_reg.fit(X_train, y_train)
         ridge_reg.score(X_test, y_test)

Out[84]: 0.7901996166492569

In [88]: ab_reg = AdaBoostRegressor(loss = "linear")
         ab_reg.fit(X_train, y_train)
         ab_reg.score(X_test, y_test)

Out[88]: 0.6924090517277011

In [89]: gb_reg = GradientBoostingRegressor(max_depth = 7, max_featu
         gb_reg.fit(X_train, y_train)
         gb_reg.score(X_test, y_test)

Out[89]: 0.6086057413153001

In [86]: dt_reg = DecisionTreeRegressor()
         dt_reg.fit(X_train, y_train)
         dt_reg.score(X_test, y_test)

Out[86]: 0.7447872870304995

In [87]: rf_reg = RandomForestRegressor()
         rf_reg.fit(X_train, y_train)
         rf_reg.score(X_test, y_test)

Out[87]: 0.8106514648362296

In [92]: xgb_reg = XGBRegressor()
         xgb_reg.fit(X, y)
         xgb_reg.score(X, y)

Out[92]: 0.9131660566803912
```

Figure 5. Accuracy

Table 2. Accuracy table

S. No.	Algorithm	Accuracy
1	Linear Regression	0.78021
2	Ridge Regression	0.79019
3	Lasso Regression	0.76216
4	Decision Tree (DT)	0.74478
5	Random Forest (RF)	0.81065
6	AdaBoost (AB)	0.69240
7	Gradient Boosting Tree (GB)	0.60806
8	Modified Extreme Boosting	0.82912

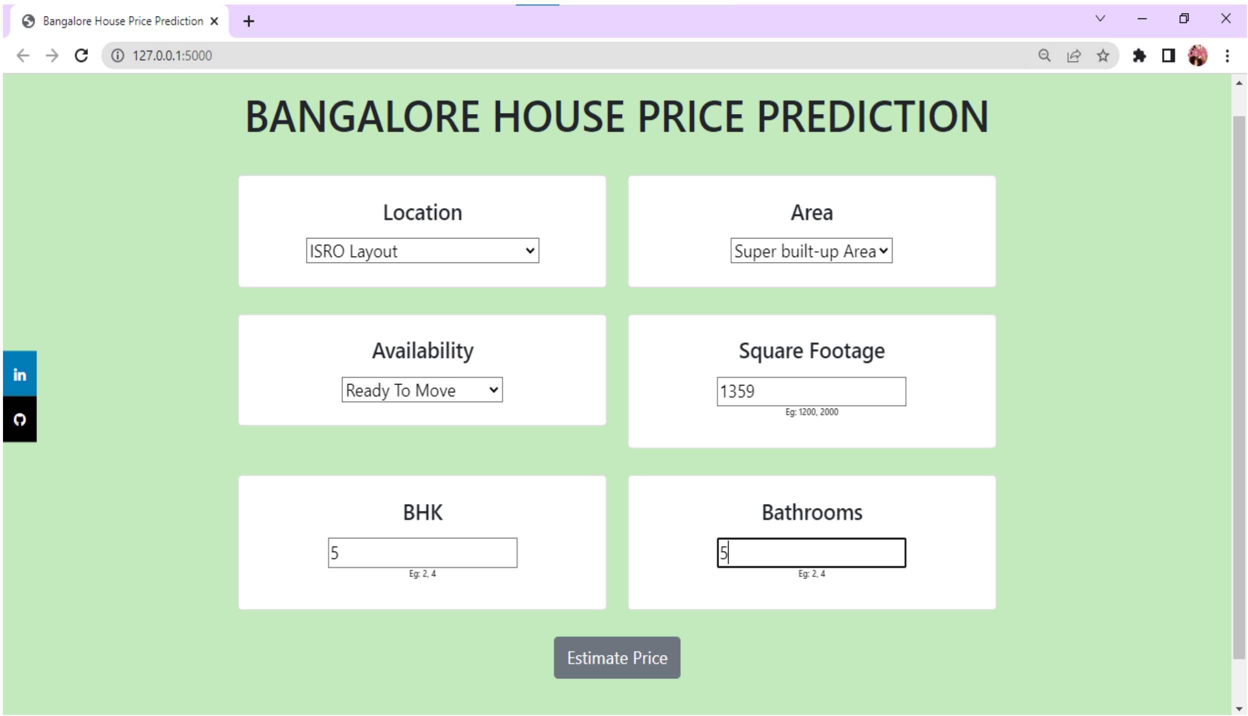
The figure 6 shows the comparison graph between the algorithms to show the performance and efficiency of the algorithms and it clearly shows that Modified XGBoost is more efficient than any other algorithms



```
In [1]: import matplotlib.pyplot as plt


        algorithms = ['Linear','Ridge','Lasso','DT','RF','AB','GB','XGBoost']
        accuracy = [0.79021, 0.79019, 0.76216, 0.74478, 0.81065, 0.69240, 0.608060, 0.91912]
        plt.plot(algorithms, accuracy)
        plt.xlabel("Algorithm")
        plt.ylabel("Accuracy")
        plt.title("Accuracy of Various Algorithms")
        plt.ylim(0, 1)
        plt.show()
```

Figure 6. Accuracy graph

This figure 7 shows that front-end of the application which is running on the server 127.0.01.5000 using flask. It contains the details such as location, area, availability, square footage, BHK, bathrooms. The input needs to be entered according the user.



BANGALORE HOUSE PRICE PREDICTION

Location: ISRO Layout

Area: Super built-up Area

Availability: Ready To Move

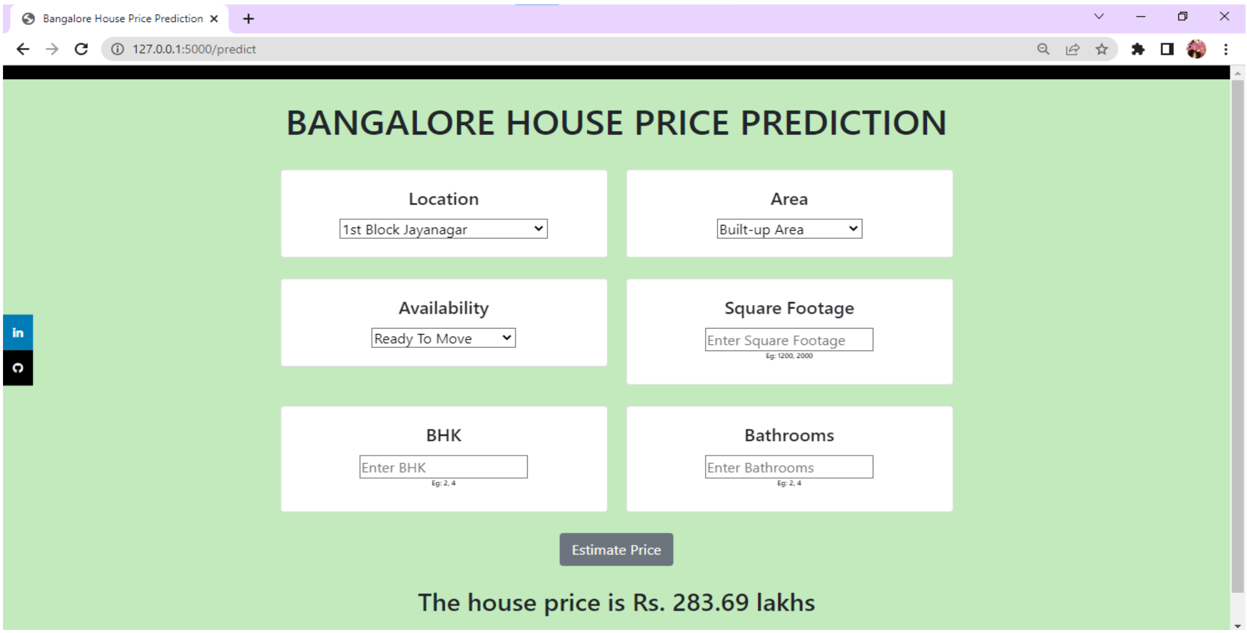
Square Footage: 1359
Eg: 1200, 2000

BHK: 5
Eg: 2, 4

Bathrooms: 5
Eg: 2, 4

Estimate Price

Figure 7. Running website on flask



BANGALORE HOUSE PRICE PREDICTION

Location: 1st Block Jayanagar

Area: Built-up Area

Availability: Ready To Move

Square Footage: Enter Square Footage
Eg: 1200, 2000

BHK: Enter BHK
Eg: 2, 4

Bathrooms: Enter Bathrooms
Eg: 2, 4

Estimate Price

The house price is Rs. 283.69 lakhs

Figure 8. House price prediction

The figure 8 shows the prediction of house price for the given input.

5. Conclusion

The creation of new jobs is a key factor in the expansion of any economy, which may be aided by the real estate sector. In this scenario, the roles of property owners and receivers are intertwined. That's why it's so important to have reliable real estate value forecasts. Property prices are a good barometer of an economy's health, so homeowners and investors pay close attention to trends in home value. Building a model that can foresee future housing costs can be a valuable tool for regulating property use and planning financially. Assisting policymakers in setting appropriate pricing and allowing property owners and brokers to make informed decisions are just two of the many benefits that come from being able to foresee a piece of real estate's future value. In this analysis, we compare the predictive efficacy of common regression algorithms to that of Bayesian Regression in predicting Bangalore home prices. Results were promising because of the abundance of features and high correlation in the publicly available data. Therefore, additional features, ideally with a high correlation to home price, need to be added to the local data. However, XGBoost produced the best results. According to the study's findings, Modified XGBoost outperforms competing prediction algorithms. Including user reviews of the property's attributes, price data from social media, images from Google Maps, and economic statistics could improve ML forecasts in the future.

Declarations

Source of Funding

This research work did not receive any grant from funding agencies in the public or not-for-profit sectors.

Competing Interests Statement

Authors have declared no competing interests.

Consent for Publication

The authors declare that they consented to the publication of this research work.

Authors' Contributions

All authors equally contributed to research and paper drafting.

References

[1] Gallent, Nick, Phoebe Stirling, and Iqbal Hamiduddin. (2022). Pandemic mobility, second homes and housing market change in a rural amenity area during COVID-19–The Brecon Beacons National Park, Wales. Progress in Planning, 100731.

[2] Gajalakshmi, R. K., et al. (2020). An Optimized ASM based Routing Algorithm for Cognitive Radio Networks. International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE.

[3] Padmini, S. V. (2022). Socio-economic conditions of slum dwellers in Karnataka-a case study in Tumkur district. Archers & Elevators Publishing House.

[4] Raj, S. Gokul, N. Srinath, and T. Ananth Kumar (2019). Real-Time Trespasser Detection Using GPS based UAV. IEEE International Conference on Innovations in Communication, Computing and Instrumentation (ICCI).

[5] Tokunaga, Robert S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in Human Behavior, 26(3): 277-287.

[6] Arunmozhiselvi, S., et al. (2022). A Systematic Approach to Agricultural Drones Using a Machine Learning Model. Machine Learning Approaches and Applications in Applied Intelligence for Healthcare Data Analytics. CRC Press, Pages 41-60.

[7] Nguyen, Quang Hung, et al. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering, 2021: 1-15.

[8] Rajmohan, R., et al. (2022). G-Sep: A Deep Learning Algorithm for Detection of Long-Term Sepsis Using Bidirectional Gated Recurrent Unit.

[9] Ford, Elizabeth, et al. (2019). Identifying undetected dementia in UK primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. BMC Medical Informatics and Decision Making, 19(1): 1-9.

[10] Muthukumarasamy, Sugumaran, et al. (2020). Machine learning in healthcare diagnosis. Blockchain and Machine Learning for E-Healthcare Sys., 343.

[11] Mair, Carolyn, et al. (2000). An investigation of machine learning based prediction systems. Journal of systems and software, 53(1): 23-29.

[12] Saravanan, V., et al. (2023). Design of deep learning model for radio resource allocation in 5G for massive IoT device. Sustainable Energy Technologies and Assessments, 56: 103054.

[13] Čeh, Marjan, et al. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS International Journal of Geo-Information, 7(5): 168.

[14] Hong, Yao, Jianhong Kang, and Ceji Fu. (2022). Rapid prediction of mine tunnel fire smoke movement with machine learning and supercomputing techniques. Fire Safety Journal, 127: 103492.

[15] Lu, Binbin, et al. (2014). Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. International Journal of Geographical Information Science, 28(4): 660-681.

[16] Alzain, Elham, et al. (2022). Application of Artificial Intelligence for Predicting Real Estate Prices: The Case of Saudi Arabia. Electronics, 11(21): 3448.

[17] Durganjali, P., and M. Vani Pujitha. (2019). House resale price prediction using classification algorithms. International Conference on Smart Structures and Systems (ICSSS), IEEE.

[18] Lu, Sifei, et al. (2017). A hybrid regression technique for house prices prediction. International Conference on Industrial Engineering and Engineering Management (IEEM), IEEE.

[19] Prasanalakshmi, B. (2022). Deep Regression hybridized Neural Network in human stress detection. International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), IEEE.

[20] Fletcher, Mike, Jean Mangan, and Emily Raeburn. (2004). Comparing hedonic models for estimating and forecasting house prices. Property Management, (3): 189-200.

[21] Hallegatte, Stéphane, et al. (2012). Investment decision making under deep uncertainty-application to climate change. World Bank Policy Research Working Paper – 6193.

[22] Sun, Shiliang, et al. (2019). A survey of optimization methods from a machine learning perspective. IEEE Transactions on Cybernetics, 50(8): 3668-3681.

[23] Coenen, Lize, et al. (2022). Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods. Journal of the Operational Research Society, 73(1): 191-206.