Predict the price of house using a Linear Regression algorithm

Ritesh Pandey
Galgotias University
Greater Noida, India
Email:
ritesh.20scse1010664@galgotiasuniversity.edu.in

Shreshth Khare
Galgotias University
Greater Noida, India
Email:
shreshth.20scse1010304@galgotiasuniversity.edu.in

Sudhanshu Tripathi
Galgotias University
Greater Noida, India
Email:
sudhanshu.20scse1010193@galgotiasuniversity.edu.in

Dr Deependra Rastogi
Galgotias University
Greater Noida, India
Email:
deependra.rastogi@galgotiasuniversity.edu.in

Abstract— Our project is based on the prediction of the price of the house using a Regression algorithm. In this project, we are using different Python libraries and machine learning. By using those libraries and datasets we will configure what can be the best price for a house. Based on geographical criteria, this approach assists in determining a base price for a property. Past market patterns and value ranges, as well as anticipated developments, can be used to forecast future costs. We will be able to create a great model for determining the price of a house by the end of this project. The main purpose of this project is to develop a faultless model that will help people make the best decision possible. We are going to perform Training and testing of the data set in CSV format as well as a data dictionary. Tools that we are going to use are Anaconda and Jupyter. It will assist clients in putting money into a bequest without having to go through a broker. The Gradient boosting regressor has an accuracy of 91.73 per cent, according to the findings of this study.

Keywords – House price prediction, Regression algorithm, Machine Learning, Gradient boosting regressor, Anaconda and Jupyter.

I. INTRODUCTION

Purchasing a home is unquestionably one of the most significant decisions one can make. The price of a house is determined by a number of factors, including the home's location, characteristics, and the demand and supply of the real estate market. The housing business is similarly important to the economy as a whole. As a result, not only buyers, but also real estate agents and economists' profit from housing value estimates. House values, growth trends, and their correlations with a number of factors are examined in housing market forecasting research. In recent years, machine learning techniques have progressed, and the explosion of data (also known as big data) has paved the way for real estate research. Many studies on the housing market have employed statistical learning approaches. The purpose of this research is to gain useful information into the housing market by analysing a genuine historical transactional dataset. It's seeking models that can predict a home's worth based on a collection of variables. Effective models may aid home purchasers and real estate brokers in making better decisions.

The goal of this study is to develop a house price prediction model that uses linear regression and a gradient boosting regressor to achieve the best possible outcomes. House price forecasts are expected to help clients who are looking to buy a home by letting them know what the price range will be in the future, allowing them to properly prepare their budget. Property investors who wish to know the trajectory of housing values in a certain location can use house price predictions. Property investors who wish to know the trajectory of housing values in a certain location can use house price predictions.

The following are the portions of the paper in order: The second segment looks at prior studies forecasting the housing market using several methods of machine learning 3rd section discusses the methodology as well as machine learning techniques. Section 4 proposes data collection and analysis. The model's In Section 5, we'll talk about implementation and evaluation, the results of this analysis will be shown in Section 6, and the conclusion and future scope of this model will be presented in Section 7.8.

II. LITERATURE SURVEY

Machine learning is a category of artificial intelligence that educates itself using readily available machinery without the need for exact coding. Machine learning is concerned with the growth of computer programmes that can adjust to new data. Machine learning algorithms can be divided into three categories: supervised machine learning, unsupervised machine learning, and reinforcement machine learning. Researchers take on historical increments rates or house price index, which are oftenly estimated from a national mean house price [3], [5], [6], or the mid point house price [4], to try to identify best ways to predict the nature of the housing market. House growth predictions, in the opinion to [1,] could serve as a leading signal for opinion-formers assessing the entire wealth. Macro factors such as income, credit, interest rates, and building costs all makes changes in house price growth. The Vector Autoregression (VAR) model was commonly used in these papers in the past [10], [11], while Dynamic Model Averaging (DMA) has gotten bigger and increasingly in-demand in past few years [3], [13], [14].

House price prediction studies, on the other face, are concerned with predicting house prices [2], [12], and [15]. These researches are in search for representation that can be used in the upcoming future to estimate the Given its features, such as location, land size, and the number of spaces, a house's price will be determined. The Support Vector Machine (SVM) and its combination with other methods are widely used to estimate the value of a home. For example, [12] combines SVM with a Genetic design to improve precision, while [15] combines SVM and Stepwise in an optimal approach to estimate house values. There are additional pre-owned methods for evaluating house values, such as cable neural networks and (PLS) [2].

METHODOLOGY ADAPTED

We are going to implement a Linear Regression algorithm in this model to predict the prices of the houses. Also, later on, we will use the Gradient boosting algorithm which is great in improving accuracy as it is too good with weak models like this.

A. *Regression*

- A Regression is a statistical analysis that attempts to predict the effect of one or more variables on another variable.
- The variable being influenced is called the dependent variable and the other variable is known as the independent variable.
- It is used in the finance or business world as an attempt to predict the effect of certain input on an output.
- For Example, an analyst of a company tries to predict the sales of its company by seeing the movements in GDP.

What is Linear Regression?

- Linear regression is a supervised learning machine learning algorithm. It carries out a regression task. Based on independent variables, regression models a goal prediction value.
- It is mostly utilised in forecasting and determining the link between variables.
- Linear Regression has one independent variable and, in most circumstances, a linear graph.

General Formula for Linear regression is:

$$\boxed{Y = a + bx + u}$$

where,

Y=The variable you are trying to predict.

X=The variable you are using to predict.

a=Intercept

b=Slope

u=The regression residual.

A residual is a measurement of how distant a point is from the regression line vertically. Simply put, it's the difference between an anticipated and actual number.

B. *Gradient Boosting Algorithm*

The primary idea behind this technique is to develop models in a sequential manner, with each model attempting to reduce the mistakes of the previous model. But how do we go about doing that? This is accomplished by constructing a new model based on the old model's errors or residuals.

One of the most powerful methods in machine learning is the gradient boosting technique. As we all know, there are two types of errors in machine learning algorithms: bias and variance errors. Gradient boosting is one of the boosting strategies that is used to reduce the model's bias error.

The diagram below shows how regression problems are solved using gradient boosted trees.

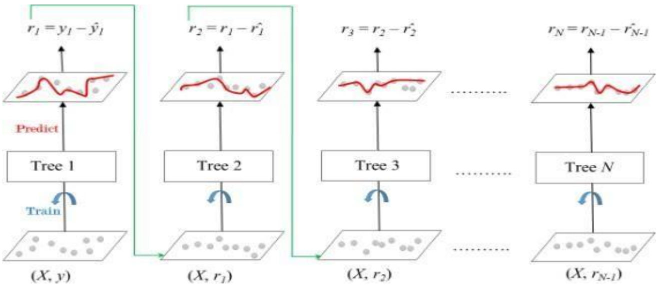
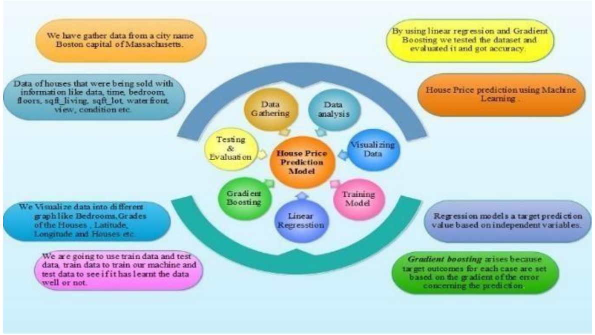


Fig. a

N trees make up the ensemble. The feature matrix X and the labels y are used to train Tree1. The training set residual errors r1 are calculated using the predictions labelled y1(hat). Tree2 is then trained using Tree1's feature matrix X and residual errors r1 as labels. Using the projected results r1, the residual r2 is determined (hat). The procedure is repeated until the ensemble's tree N has been trained completely.

C. *Block Diagram*

The Block Diagram given below is the summary of the methodology explained in the paper below-



We have gather data from a city name Boston capital of Massachusetts.

Data of houses that were being sold with information like data, time, bedroom, floors, sqft_living, sqft_lot, waterfront, view, condition etc.

We Visualize data into different graph like Bedrooms,Grades of the Houses , Latitude, Longitude and Houses etc.

We are going to use train data and test data, train data to train our machine and test data to see if it has learnt the data well or not.

By using linear regression and Gradient Boosting we tested the dataset and evaluated it and got accuracy.

House Price prediction using Machine Learning .

Regression models a target prediction value based on independent variables.

Gradient boosting arises because target outcomes for each case are set based on the gradient of the error concerning the prediction.

III. DATA GATHERING AND ANALYSIS

The data used in the analysis is a survey report of Boston city, capital of Massachusetts in which data of houses that were being sold with information like date, time, bedroom, floors, Sqft_living, Sqft_Lot, waterfront, view, condition and many more. The Data description is given below:-

Data Used	Description
Date	Date when house was sold
Price	Price targeted for prediction
Bedrooms	Number of Bedrooms per House
Bathrooms	Number of bathrooms per House
Sqft_living	Square footage of the home
Sqft_Lot	Square footage of the lot
Floors	Total floors in the house
Waterfront	House which has a view of a waterfront
View	Viewed
Condition	Overall condition
Grade	The grade(given to the housing unit)
Sqft_Above	Square footage of house apart from the basement.
Sqft_Basement	Square footage of the basement
yr_Built	Built in Year
yr_Renovated	Year when the house was renovated
Lat	Latitudes coordinate
Long	Longitudes coordinate
Sqft_Living20	Living room area in 2020
Sqft_Lot20	Lot Size area in 2020

A. Data Visualization

1) Bedrooms vs Number of Houses

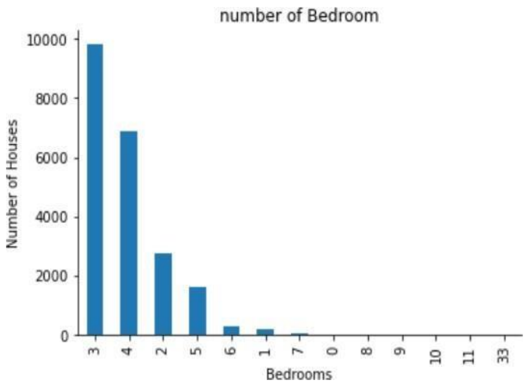


Fig .1

Fig. 1 shows the graph between several houses and the number of bedrooms. From this graph, we can say that the number of houses is more in which there are 3 bedrooms.

2) Grades of the Houses vs Number of Houses

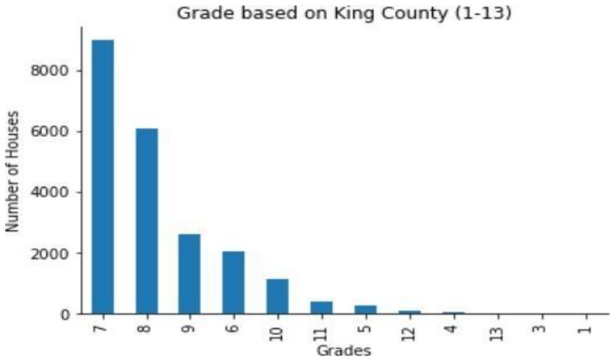


Fig .2

Fig. 2 shows the number of houses vs the grade graph. From this graph, we can say that the number of houses is more which are graded as 7 according to the king county grade.

3) Bedrooms vs Number of Houses

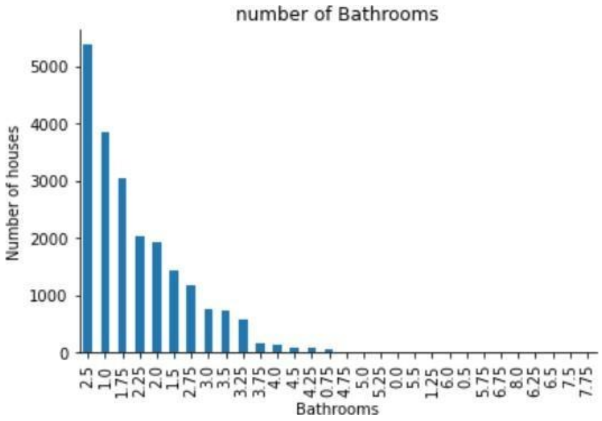


Fig. 3

Fig. 3 shows the number of houses vs several bathrooms graph. From this graph, we can say that the number of houses is more in which there are 2 to 3 bathrooms.

4) Visualizing the Latitude, Longitude and Houses

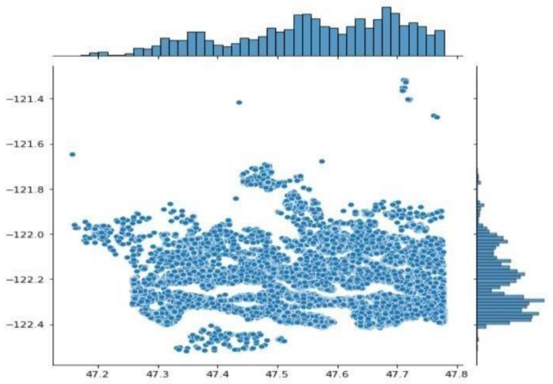


Fig.4

On the dataset, the x-axis represents latitude and the y-axis represents longitude for each house. We can notice the general placement and layout of the houses. We use seaborn and get the visual shown above. The join plot tool can be incredibly useful in determining where information is concentrated and where it is placed. Let's examine what we can learn from this representation. There are a lot of houses between -47.7 and -48.8, which could indicate that it's a good area for buyers.

5) *Heat Map*

Now, we will visualize the common factors which are going to affect the price of the house. But before that, we are going to plot a Heatmap that shows all the price affecting factors in one view.

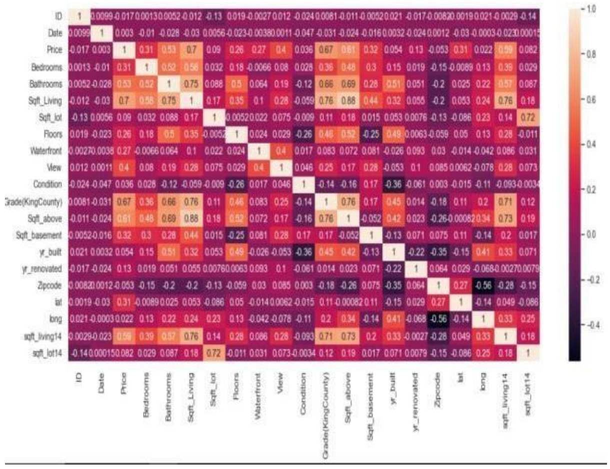


Fig. 5

From the above figure that is representing Heatmap, we can see that some common factors such as Sqft_living, Sqft_above, Grades, Bedrooms, Bathrooms, and Location of the houses are some factors that will be going to affect the prices of the houses.

A heatmap is a two-dimensional graphical representation of data that uses colours to represent the individual values in a matrix.

Users can produce annotated heatmaps with the seaborn python module, which can be altered using Matplotlib capabilities to match their needs.

6) *Price vs Square feet graph*

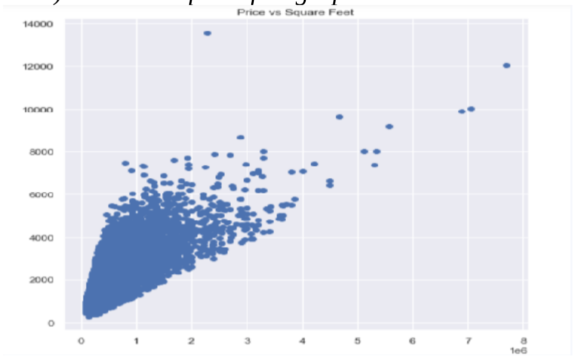


Fig. 6

A scatter plot is the type of graph we used above. A scatter plot is a type of graph that shows how our data points are distributed and is commonly used for two variables. The first graph shows that the larger the living area, the higher the price, even if the data is concentrated in one price range.

7) *Price vs location*

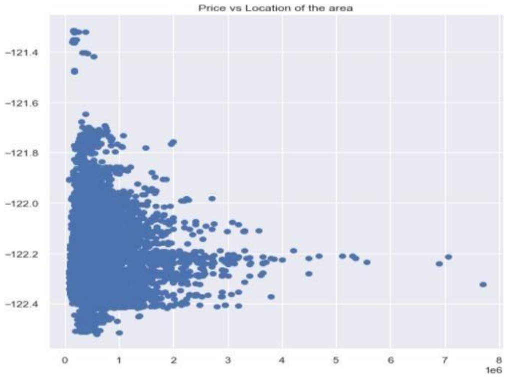


Fig..7

This graph shows where the houses are located in terms of longitude, and it makes an intriguing observation: -122.2 to -122.4 sells houses for a substantially higher price.

Now, we will see some more factors affecting the prices of the houses.

8) *Bedroom vs Price graph*

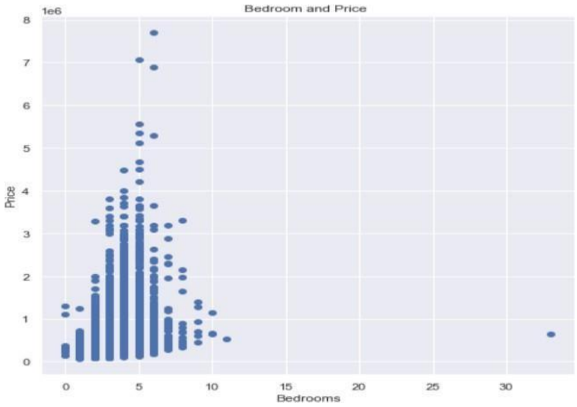


Fig. 8

The above graph shows us the prices of the houses according to the bedrooms. After analysing this we can say that the prices of the houses are more in where there are 3 to 5 bedrooms.

9) *Sqft living and Sqft basement vs Price graph*

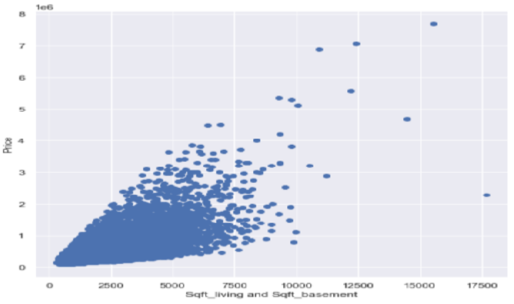
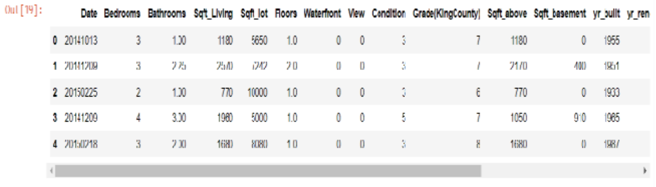


Fig.9

The above graph tells us that the more the Sqft living area, the more the prices of the houses are.

After visualizing all the important and common factors that are going to affect the price, we can say that the ID and Date of the houses don't have much effect on the prices of the houses.

So, we have dropped the date and ID from our training Dataset such that it will be simple to train the dataset.



Out [19]:

	Date	Bedrooms	Bathrooms	Sqft_Living	Sqft_lot	Floors	Waterfront	View	Condition	Grade(KingCounty)	Sqft_above	Sqft_basement	yr_built	yr_ren
0	20141013	3	1.00	1180	5650	1.0	0	0	3	7	1180	0	1955	
1	20141209	3	2.25	2570	7242	2.0	0	0	3	7	2170	400	1951	
2	20150225	2	1.00	770	10000	1.0	0	0	3	6	770	0	1933	
3	20141209	4	3.00	1960	5000	1.0	0	0	5	7	1050	910	1965	
4	20150218	3	2.00	1680	8080	1.0	0	0	3	8	1680	0	1987	

Fig.10

After dropping the unimportant ID and Date from the dataset, we are going to visualize some more factors which are going to influence the prices of the houses.

10) House vs Floors graph

From the graph below we can analyse that the number of houses is more in which there is only 1 floor.

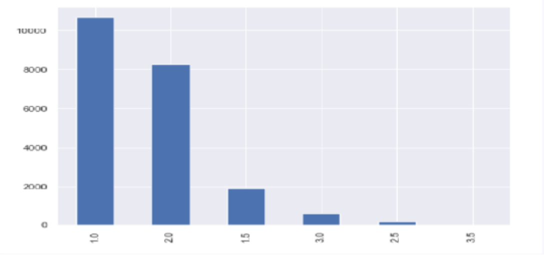


Fig. 11

11) Price vs Zipcode

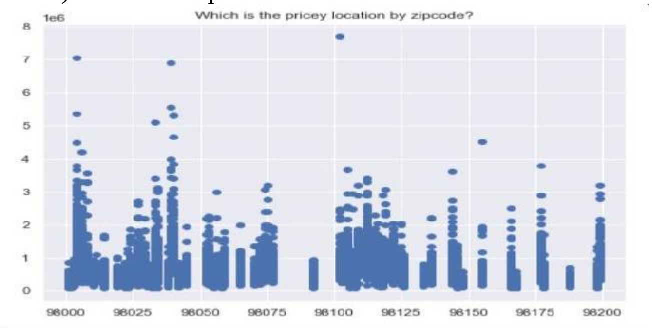


Fig. 12

In this comparison, we have shown how the price of a house fluctuates with the Zipcode. The posher area the more the price will be.

IV. TRAINING THE MODEL

Train data will be used to train our machine, and test data will be used to verify if it has learned the data correctly.

The criterion variable, often known as Y, is the variable we're trying to forecast. The variable on which our predictions are based is known as the predictor variable and is abbreviated as X.

TESTING AND EVALUATING THE MODEL BY LINEAR REGRESSION ALGORITHM

Firstly, we have to use the sci-kit to learn a model for importing the linear regression algorithm then all the steps are as follows.

Step by Step Explanation :

1) As previously stated, we will use linear regression for our model, and to do so, we will utilise scikit learn (a built-in python library) and import linear regression from it.

2) Linear Regression is then set to a variable reg.

3) Because our primary purpose is to anticipate prices, we set labels (output) as price columns and convert dates to 1s and 0s to minimise the impact on our data. We use 0 for houses that were built after 2014 and are brand new.

4) From Python's sci-kit learn module, we import train test split to split our data into test and train.

5) We'll train 90% of the data and test 10% of it, and we'll use random state to randomised the data splitting.

6) Now that we have both train and test data, as well as labels for each, let's fit our train and test data into the linear regression model.

7) We have a 73 percent accuracy after fitting the data to our model. Our expectations were not met.

So now we'll apply a different strategy that'll be quite useful in a model like this with a weak prediction model.

The Gradient Boosting approach will be employed.

A. TESTING AND EVALUATING THE MODEL BY GRADIENT BOOSTING ALGORITHM

Gradient boosting is a type of machine learning enhancement. It is based on the assumption that when the best potential next model is coupled with prior models, the overall prediction error is minimised. To reduce the error, the fundamental concept is to specify the desired outcomes for the following model.

Gradient boosting gets its name from the fact that the goal outcomes for each case are determined by the gradient of the prediction error. In the set of feasible predictions for each training case, each new model makes a step in the direction that minimises prediction error.

Let's look at the steps,

First of all, we import the ensemble from the sci-kit learn library.

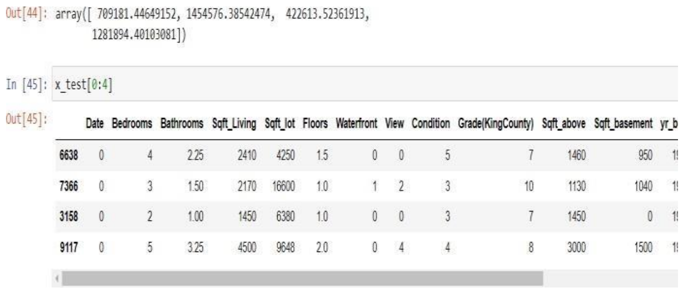
Then, for our gradient boosting regressor, we create a variable and set some parameters for it, as seen above.

- n estimator — The number of boosting phases that must be completed. We shouldn't set it too high because that will make our model look too big.
- max depth — The tree node's maximum depth.
- learning rate — The rate at which the data is learned.

- loss — to be optimised loss function The least squares regression is denoted by the letter 'ls.'

V. RESULT

We checked for accuracy scores after fitting our training data into the gradient boosting model, and we received an accuracy of **91.85** per cent which is way better than the accuracy we got after testing the linear regression model.



```
Out[44]: array([ 709181.44649152, 1454576.38542474,  422613.52361913,
                1281894.40103081])

In [45]: x_test[0:4]

Out[45]:
        Date  Bedrooms  Bathrooms  Sqft_Living  Sqft_lot  Floors  Waterfront  View  Condition  Grade(KingCounty)  Sqft_above  Sqft_basement  yr_b
  6638     0         4       2.25         2410      4250     1.5           0     0          5                  7        1460            950     1
  7366     0         3       1.50         2170     16600     1.0           1     2          3                 10        1130           1040     1
  3158     0         2       1.00         1450      6380     1.0           0     0          3                  7        1450              0     1
  9117     0         5       3.25         4500      9648     2.0           0     4          4                  8        3000           1500     1
```

Fig. 13

In the above figure, we can see that our prediction model is working properly as it predicted the first four house's prices.

For example, a house with **4** bedrooms, **2** normal size bathrooms and **1** small bathroom and has **2410** square feet of living area has a price of **709181.44** Rupees (Any Suitable Currency).

Algorithm	Accuracy
Linear Regression Algorithm	73.20%
Gradient Boosting Algorithm	91.85%

Fig.14

In fig.16 we can see our algorithms and their accuracy.

Now that our model is up and running, we want to calculate errors like

1. Mean Squared Error.

2. Absolute Mean Error

3. Root Mean Squared Error

But before that, we have plotted a scatter graph to check that our model is working properly or not in a graphical way.

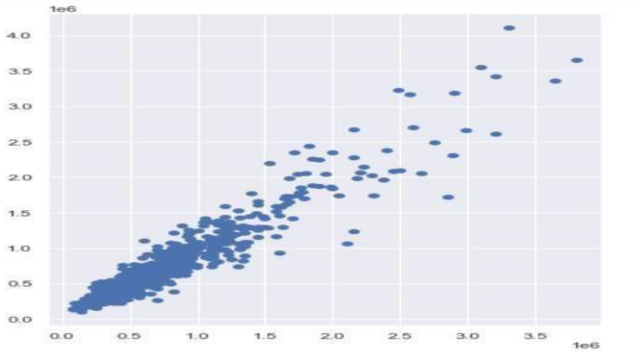


Fig.15

From the above graph, we can conclude that our graph is **linear** and as we know the graph of linear regression is linear. So now, we can say that our model is working properly.

Now, we have imported metrics from the sklearn library to calculate our errors.

Error found in the Data	Value
Mean Absolute Error	9578.230
Mean Squared Error	259163.482
Root Mean Squared Error	1081.943

Fig.16

From the above figure, we can see that the errors are minimized considering our big and complex dataset.

VI. FUTURE SCOPE

We will provide a comparative study of the system's predicted price and the price from real estate websites such as Housing.com, no broker. in, and 99acers.com in the future for the same user input. To make things easy for the consumer, we'll additionally recommend real estate properties based on the predicted pricing. The dataset now only includes cities in Boston; however, it is planned to expand to other Indian cities and states in the future. To make the system even more informative and user-friendly, we'll also use a G map to make the system more effective. This will display the nearby amenities, such as hospitals and schools, within a large area of the given location. This can also be considered while making predictions, as the presence of such elements can influence the outcome. To improve the outcome, even more, future work on this topic could be separated into numerous areas. This can be accomplished by:

1. The applied pre-processing procedures aid in the accuracy of prediction. Experimenting with different combinations of pre-processing procedures to improve forecast accuracy is a good idea.
2. Utilize the various features and see if they can be blended as binning features to improve the data.

The additional feature that can be added to our suggested system is to provide users with a full-featured user interface, allowing them to use the ML model for many locations with multiple functionalities.

VII. CONCLUSION

In this research, a prediction model is built using Linear regression and Gradient boosting machine learning algorithms to forecast possible selling prices for any house property. Additional information such as longitude and a heat map was added to the dataset to aid in price prediction. These characteristics are not commonly found in other prediction systems' datasets, which distinguishes this system. These features influence people’s decisions while purchasing a property, so why not include them in predicting house prices. The system provides 91.85% accuracy while predicting the prices for the house prices which is way far better than the model out there in the market.

REFERENCES

[1] R. Gupta, A. Kabundi, and S. M. Miller (2011). Structural and non-structural models with and without fundamentals are used to forecast the US real house price index. Economic Modelling, vol. 28, no. 4, 2013, pp. 2013-2021.

[2] Housing Value Forecasting Using Machine Learning Methods, Mu, J., Wu, F., and Zhang, A., 2014. 2014(2014), p.7, Abstract and Applied Analysis.

[3] Bork, L., and Moller, S., 2015. Forecasting housing prices in the 50 states with Dynamic Model Averaging and Dynamic Model Selection. 63–78 in International Journal of Forecasting, 31(1).

[4] Balcilar, M., Gupta, R., and Miller, S. M. Balcilar, M., Gupta, R., and Miller, S. M. (2015). Nonlinear models of regional housing prices' out-of-sample forecasting performance in the United States. Applied Economics, vol. 47, no. 22, pp. 2259-2277.

[5] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928-2934.

[6] Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the US real house price index. Economic Modelling, 45, 259-267.

[7] Ng, A., & Deisenroth, M. (2015). Machine learning for London housing price prediction mobile application. Technical Report, June 2015, Imperial College, London, UK.

[8] C. Rahal, C. Rahal, C. Rahal, C. Rahal, C. Rah (2015). Forecasting House Prices Using Factor Combinations (No. 15-05). M. Risse and M. Kern, 2016. Dynamic model averaging is used to forecast house price growth in the Eurozone. pp. 70–85 in North American Journal of Economics and Finance, vol. 38, no.

[9] T. J. Z. Jie, T. J. Z. Jie, T. J. Z. Jie, T. J. (2005). Evidence from Shanghai on what drives up property prices. [J]. 5, 005 (World Economy).

[10]X. K. M. Y. D. Changrong, X. K. M. Y. D. Changrong, X. K. M. (2010). China House Price Volatility Clustering and Short-Term Forecast [J]. The Chinese Journal of Management, vol. 6, no. 024, is a publication dedicated to the study of management.

[11]J. Gu, M. Zhu, and L. Jiang, 2011. A genetic algorithm and a support vector machine are used to forecast housing prices. pp. 3383–3386 in Expert Systems with Applications, vol. 38, no. 4.

[12] Wei Y. and Cao Y., 2017. House price forecasting using the dynamic model averaging approach: Evidence from China. 147–155 in Economic Modelling, vol. 61.

[13]P. F. Chen, M. S. Chien, and C. C. Lee (2011). In Taiwan, dynamic modelling of regional housing price dissemination. The Journal of Housing Economics, vol. 20, no. 4, pp. 315-332.

[14] J.-H. Chen et al., 2017. Support Vector Machine is used to forecast the geographic dynamics of the housing market. pp.273–283 in International Journal of Strategic Property Management, vol. 21, no. 3.