GSJ: Volume 11, Issue 11, November 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

# House Price Prediction using OLS and XGBoost Regression

Zubair Ahmad

Department of Computer Science & Technology

University of Sialkot

Sialkot, Pakistan 21101009-018@uskt.edu.pk

***Abstract – It is very important to know about what's happening around us in every field of life. Machine Learning made it obviously easy for us to do it without wasting our so much time. Product analysis, Image processing, Pattern recognition, medical diagnosis, and so on are many fields where machine learning is directly employed. Mainly here we are discussing machine learning implications in the field of real estate to predict the prices of house. So, many factors are taken into consideration to predict the house price like, bed rooms, TV lounge, decoration style, balcony, garage, parks, nearly hospitals, and broad roads and so on.***

***We will surely predict house price on the base of different metrics by combining different techniques of machine learning. Combination of OLS Regression, Linear regression, random forest and XGBoost regression will give us a very concise result.***

***Keyword:*** *OLS Regression, Linear Regression, XGBoost Regression, House Price Prediction*

## 1. Introduction

There are so many machine learning implication in every field of life either the field is arts based or science based. Health industries, production companies, ticketing, spam emails, pattern recognition, speech recognition, image processing and many more are different applications of machine learning. Most of the time real world problems are very complicated due to so many constraints related to various attributes of data [1]. Long ago, when people decide to sell or purchase property face a lot of problems and got an error of 25% which leads them to loss a very big amount of money. Machine learning replaces previous technologies with modern technology. It implement the latest trends in real estates. All its working depends upon the data. Sometime dataset is called backbone of any field in the machine learning project because machine trains itself automatically by using datasets of same and sometime of different kind which helps to predict new data on the basics of previous data. As population is increasingly day by day with higher rate and people from rural areas due to less opportunities of jobs migrating to the cities so demand for housing is also increasing rapidly. So, people who don't now the actual price and market trends suffer a lot in sense of money. Training dataset of 80% and testing dataset of 20% would be used for the implementation of this project which would help to find the best accuracy rate to minimization of the error rate [2]. There are many people that always think to invest their money to save it for future and when they invest in especially real estate they face many hurdles due to some factors and they try to find some question's answers that is this the right place for investing the money? Is it okay to investing here? Am I paying the right price or not? So, by taking into consideration multiple factors of real estate we have discussed here the accurate and right answer of these question. XG boost is the ensemble learning technique that ensemble multiple models at a single place in numerical number with some ratios to training and testing data and gives the right answers [1].

## 2. Literature Review

In this survey for the acquiring of best results author Vargas and Silva conduct a comparative study about the house prices and showed that price of the house is the main pillar in the business of real estate. And with accumulating other factors like developmental work and facilities regarding that

increases the house price with greater rate. By focusing these factors at the time of selling the house, householders are pressurized to reduce the price. So, cost of house strictly fall due to this bargain and people get sometime big loss of money.

The actual cost of the house also depend on the land and here we get by using some other set of factors with different numerical methods. As numerical numbers are easy to memorize and when actual information come with some numbers we gain additional clearance. As house prices are increasing every year so, there would be a somehow best way to predict the price by taking some factor for relationship between dependent and some independent which would help the developer to foresee his investment and client get the genuine cost if he make himself to buy the house. [3]

House price prediction is difficult in a common way because when we contact with even real estate manager we still face a lot of problems and do not satisfy with that as we are on the great risk of loss. The whole investment can go wrong if a person cross over and leading to great loss. So keeping in view that factors which are truly effects the cost price of the house a recently study has been deployed as the key kernel for Kaggle Challenge "House Prices: Advanced Regression Technique"[4].

## 3. Data Collection about its Problem

For every research data is very important and this data collection of real estate was done from Ranchi capital of Jharkhand where it was dependent on various factors like no of bedrooms, its locations, it's number of drawing rooms, material which was used in the house, interior of the house, it's parking area, its area per square feet and it's lot area [3]. By keeping in mind these factors the data gathered about real estate from its, zone, sub divisional area of Ranchi, Jharkhand. That are shown and labeled [3].

## 4. Techniques of Machine Learning for building model

TABLE 1- PRICE FACRTORS

| Avg. Price / Sqft | area | lotsize | bedrooms | bathroom |
|---|---|---|---|---|
| ₹2,371 | rnc | 850 | 3 | 1 |
| ₹3,761 | rnc | 4000 | 4 | 2 |
| ₹3,724 | rnc | 3060 | 3 | 1 |
| ₹3,724 | rnc | 6650 | 3 | 1 |
| ₹3,097 | rnc | 6360 | 3 | 2 |
| ₹4,093 | rnc | 7383 | 6 | 3 |
| ₹3,396 | rnc | 6734 | 5 | 2 |
| ₹3,396 | rnc | 9866 | 4 | 1 |
| ₹3,396 | rnc | 7888 | 4 | 1 |

© GSJ

The actual methods used by author [3] is OLS regression technique in linear regression where it was determine that how the price of the house is increasing on depending some factors.

OLS Regression

OLS is stands for Ordinary least square method that is firstly estimate the unknown parameters in dataset and minimize the sum of squared errors in known results of data and predicted results of the data. Straight line shows the results on the right hand side of the graph in each case and on left side to vertically predicted factors and on the horizontally there are observed factors are placed [3]. Linear Regression is very helpful to check the results very easily that are going predicted on some factors of dependent and independent. It helps out the organization or investor to foresee its investment results before they invest.

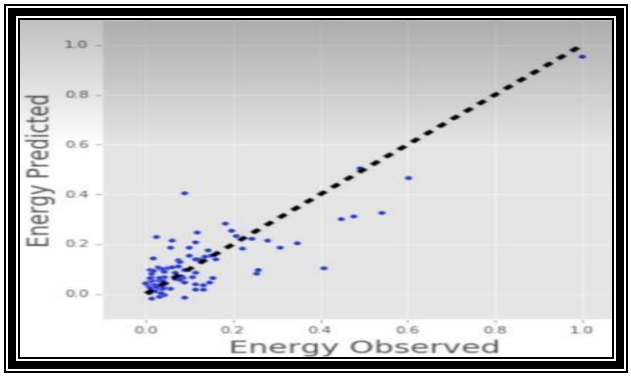



Figure 1- OLS regression [3]

TABLE 2 - COST DRIVERS [3]

| no of floors | driveways | fullbaseme | garage | balcony |
|---|---|---|---|---|
| 2 | yes | yes | yes | no |
| 2 | yes | no | yes | no |
| 1 | no | yes | yes | yes |
| 2 | no | no | yes | yes |
| 3 | no | yes | yes | no |
| 2 | yes | no | no | yes |
| 1 | yes | yes | no | no |
| 3 | yes | no | no | yes |
| 2 | ys | yes | Yes | no |

© GSJ

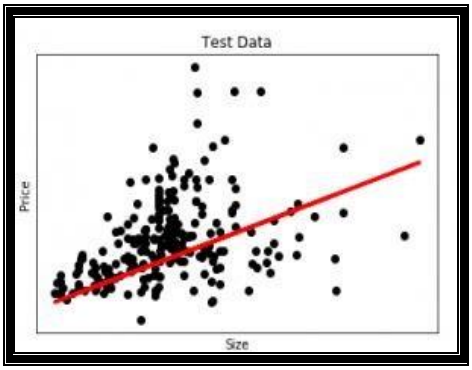



**Figure 2- test data set of real estate [3]**

## 5. Methodology and Results

As discussed earlier that dataset is gathered from different zone and locations of Ranchi zone. Feature datasets are the metrics or features to predict the house price. This approach comprises upon 19 different metrics or features to predict the house price.

Feature extraction is vital to prediction using machine learning and finding the coefficient correlation value that is between some ranges for strong from 0.3 to 0.5 and weak level is between 0.2 to o.29 and very weak is between to 0.1 to 1. [3]

Features table will shows the feature of datasets for calculating the real estate price truly depend upon.

| Features | Description |
|---|---|
| Selling price | Dispose price/sqf (RM) |
| Buying price | Transaction price/sqf (RM) |
| Floor | Floor |
| GC | Green certificate |
| MFA | Main floor area |
| Bed | Number of bedrooms |
| Distance | Distance to CBD |
| BC | Building category |
| Ownership | Own |
| CA | Category area |
| AC | Area classification |
| Floor | Floor |
| BC | Building category |
| CLASS | Building classification |
| Bed | Number of bedroom |
| Age | Age of the building |
| Buy | Buyers |
| Sell | Seller |

Figure and table 4- features

Table 5-OLS regression result [3]



OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared (uncentered): | 0.956 |
| Model: | OLS | Adj. R-squared (uncentered): | 0.956 |
| Method: | Least Squares | F-statistic: | 1067. |
| Date: | Mon, 15 Jul 2019 | Prob (F-statistic): | 0.00 |
| Time: | 06:03:17 | Log-Likelihood: | -6034.8 |
| No. Observations: | 546 | AIC: | 1.209e+04 |
| Df Residuals: | 535 | BIC: | 1.214e+04 |
| Df Model: | 11 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Lotsize | 3.4431 | 0.339 | 10.144 | 0.000 | 2.776 | 4.110 |
| bedrooms | 1095.9263 | 842.938 | 1.300 | 0.194 | -559.947 | 2751.800 |
| bathrms | 1.402e+04 | 1466.301 | 9.561 | 0.000 | 1.11e+04 | 1.69e+04 |
| stories | 6526.5732 | 925.283 | 7.054 | 0.000 | 4708.940 | 8344.206 |
| driveway | 5665.6447 | 1854.971 | 3.054 | 0.002 | 2021.724 | 9309.565 |
| recroom | 4659.4642 | 1896.548 | 2.457 | 0.014 | 933.870 | 8385.059 |
| fullbase | 5306.1054 | 1583.810 | 3.350 | 0.001 | 2194.856 | 8417.355 |
| gashw | 1.285e+04 | 3218.757 | 3.993 | 0.000 | 6529.985 | 1.92e+04 |
| airco | 1.28e+04 | 1549.330 | 8.260 | 0.000 | 9754.655 | 1.58e+04 |
| garagepl | 4379.7318 | 833.106 | 5.257 | 0.000 | 2743.173 | 6016.291 |
| prefarea | 9561.2358 | 1661.849 | 5.753 | 0.000 | 6296.687 | 1.28e+04 |

| | | | |
|---|---|---|---|
| Omnibus: | 101.942 | Durbin-Watson: | 1.576 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 279.382 |
| Skew: | 0.915 | Prob(JB): | 2.15e-61 |
| Kurtosis: | 5.988 | Cond. No. | 2.74e+04 |

Linear Regression uses random ordinary smallest square technique to the datasets of real estate different 19 features to refine the datasets metrics. Which tells that change in the dependent variable "house price" resulting from one unit change in particular metrics, but all other being held constant. So, result would be determined that a house having maximum facilities will charge maximum and would be highly influenced.

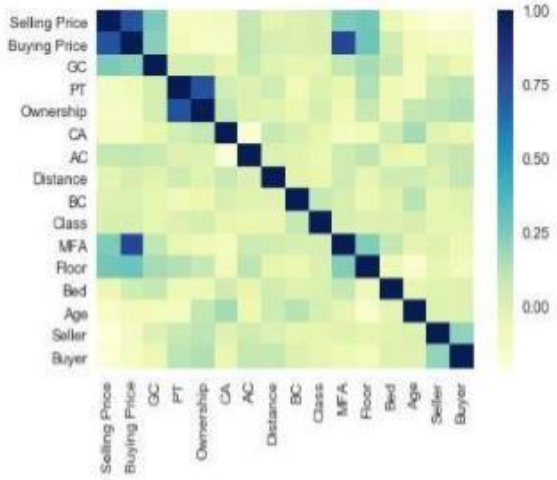
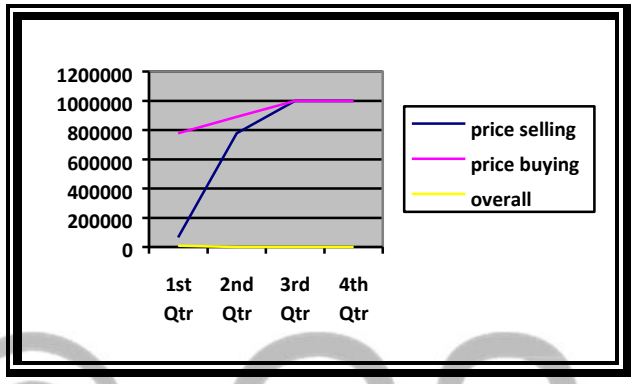



Figure 5- Correlation level of all features


**Figure 6 - Correlation level of features with values**


**Figure 7 - Correlation level of features with values between (0.1 to 0.19)**


**Figure 8 – The Correlation level between selling and buying price [3]**




**Fgure 9 – Prediction in increasing slope of price[3]**

Before giving data into the model for next working the data pre-processing is very important process for refining the data. This process has four stages that are once fulfill then the process goes with flow and result can be maximized as house price truly depends upon the data. Its stages are i) Data cleaning ii)Data Editing iii) Data reduction iv) Data wrangling. In Data cleaning no data field can be empty filled or inaccurate. If it happens in the data then it can filled using mean or median process or the whole record deleted from the data for the better results. If there is a data that is not very important then it is eradicated as it can disturb the accuracy of the result. Z score process is data reduction process which works like a normalization in data reduction which helps out for the better results. Production of graphs in filtration of the data before going to the model for acquiring results these terms comes under the banner of data wrangling [1]

As most of the time data is in the categorical form and by using one hot encoding it converted into the number form because for algorithmic process it is very important to convert into this form. As preferred conversion of this form is binary because it is very easy to mapping the values where data values represented by 1 and others denoted by the 0. For Example, in categorical manner colors having different values red, red, yellow, green but in the binary representation it is termed as

[1,0,0], [1,0], [0,1,0], [0,0,1] and when this data would be fit in model it would be easy to understand for machine learning algorithms.

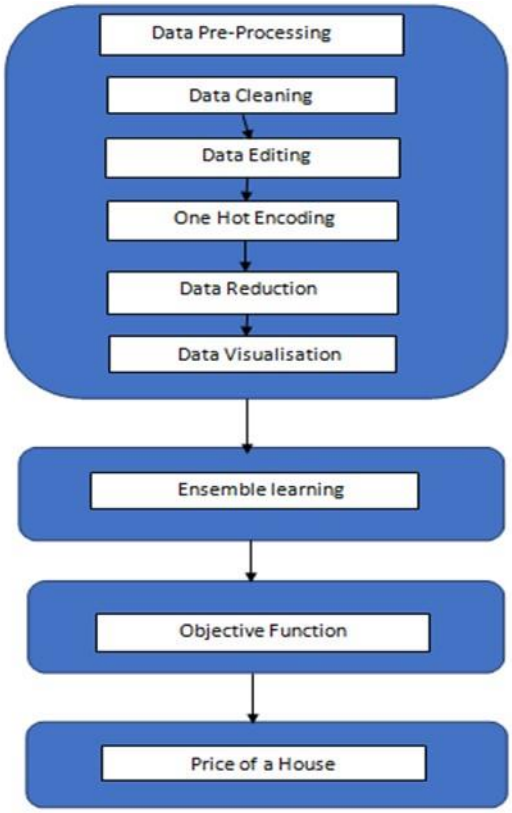



Figure 10- Flow Diagram for house price prediction [1]

For getting the results from different models and ensemble them author [1] use XGBoost regression which best supervised learning algorithms consisting objective function which includes loss function that encompasses the difference between actual and predicted value even how far it away from actual value and base learner in machine learning especially in ensemble learning XGBoost uses many models as base learners which are consider to predicting single value. Base learner also termed as weak learner and by summing up all weak learner (bad prediction) cancelled out by strong learner (good prediction).

5.1 Here are some steps involve in the XG Boost Regression Algorithm for House Price Prediction

Input: Data set of house price metrics

Output: Price of House

1. Check input dataset for missing values and calculate the mean is replaced in place of missing value.
2. Divide attributes based on values in data fields as categorical and non-categorical rows.
3. Check non categorical rows for outliers using outlier detection techniques and remove all outliers.
4. Convert categorical rows into binary vectors using one hot encoding.
5. Divide dataset for cross validation using train test split.
6. Apply Ensemble learning through training and combining individual models termed as base learners in order to derive a single prediction.

a) Calculate Mean Squared Error (MSE) with true vales to predict.
b) Classify independent models as weak-learners and strong-learners using error detection.
c) Total mean cancels bad prediction with good prediction.

7. Objective function contains the loss function and regularization term to calculate difference between actual and predicted value.

Here is plot graph of training dataset of 80 attributes having nearly 1500 records from Kaggle competition on house price prediction [1] implementation in Anaconda software will look like.

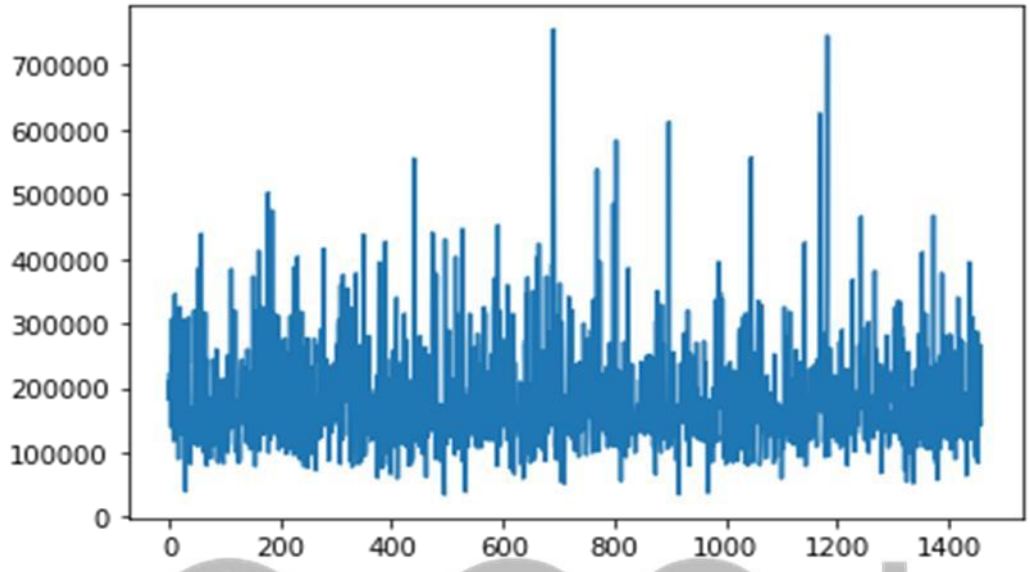



Fig 11- Plot graph for Train Dataset

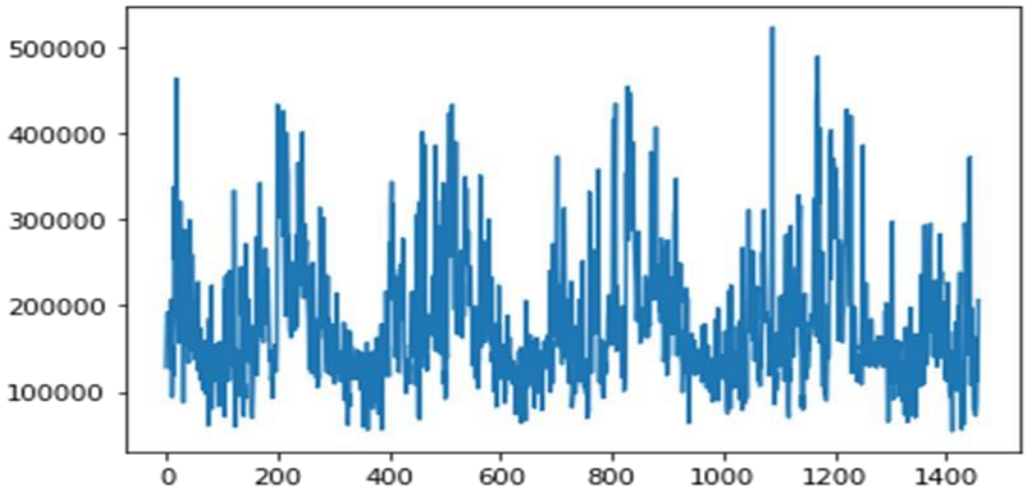
After plot graph of train dataset three different models of training, validation and testing to find test error used and get the appropriate result at different proportions of these three datasets.

Table.3. Mean Test Errors for 3 Different Models on Different Runs [1]

| Training Data (%) | Testing Data (%) | Validation Data (%) | Test Error (%) |
|---|---|---|---|
| 60 | 20 | 20 | 5.4 |
| 70 | 15 | 15 | 4.6 |
| 80 | 10 | 10 | 4.4 |
| 90 | 5 | 5 | 4.8 |

When these three different models are used with different dataset proportion then results would be different on each runs. By changing their values we will get the best results with low test Error rate. As from the table dataset training 80, testing 10, validation 10 and test error at 4.4 is best results. As low test error rate resulting into higher accuracy rate. So, it's

plot graph will look like.




## 6. Conclusion

From the dataset which we author [3] has collected from online source of Ranchi zone of Jharkhand to predict the price of the house that predict the price of house depending upon the 19 feature or metrics.

Machine Learning is widely helping in every field of life but especially in real estate which discussed here that how smoothly a house can be predicted by machine by following some metrics and can give very accurate results. [3]

XGBoost regression algorithm gives very outstanding results and by focusing on results buyer can easily buy property and seller can sell property. Number of feature can also increase the accuracy of the result. It helps to satisfy the needs of the customer and decreasing the risk of loss investing into the real estate [1].

We conclude the results after implementation of **OLS Regression, linear Regression** and **XGBoost Regression** that we get good results from both of them with higher accuracy rate. And we can achieve more results by adding both of them in the future to predict the house price with higher accuracy rate.

## 7. Referencing

1. Jangaraj, Avanija & Sunitha, Gurram & Madhavi, Reddy & Kora, Padmavathi & Hitesh, R & Associate, Sai. (2021). Prediction of House Price Using XGBoost Regression Algorithm. Turkish Journal of Computer and Mathematics Education (TURCOMAT). 12. 2151-2155.
2. RAWOOL, ANAND G., DATTATRAY V. ROGYE, and SAINATH G. RANE. "House Price Prediction Using Machine Learning." (2021).
3. Sinha, Anurag. "Utilization Of Machine Learning Models In Real Estate House Price Prediction."
4. Quang Troung, Minh Nguyen, Hy Dang, Bo Mei. House Price Prediction via Improved Machine Learning Techniques. Precedia Engineering (2020); 174:433-442.