International Journal of Technology, Management and Social Sciences (IJTMSS)
Volume 1_Issue 1 [October-December 2022]
www.ijtmss.org

PREDICTION OF HOUSING PRICES OF REAL ESTATE BUSINESS THROUGH MACHINE LEARNING

Pratyakshi Sarma[1] and Bapan Kalita[2*]

[1,2]Department of Mathematics, The Assam Royal Global University, Assam, India

Abstract

Now a day, people prefer purchasing house from real estate agencies. Different agencies have different price lists. Customers as well as the agencies find it difficult to fix the suitable price of a house. Machine learning has been playing a major role from past years in image detection, spam reorganization, normal speech command, product recommendation and medical diagnosis. In our study, we discuss the prediction of housing prices of real estate business that is generated by machine learning algorithm. After checking the multicollinearity, multiple linear regression method has been incorporated to the data as the machine learning algorithm. Comparing the predicted prices of these cities it is observed that Mumbai tops the price list followed by Delhi, Chennai, Kolkata, Hyderabad accordingly and Bangalore ends the price list.

Keywords: *Real estate; Price Fixation; Machine Learning; Linear Regression; Supervised Learning*

1 Introduction:

1.1 Real Estate

Real estate requirement is expected to increase by 17 million square ft. by 2025. Demand for household assets has surged because of rising urbanization and increasing house-rent income. Indian real estate market has already occupied in the leading 10 international markets. Real estate is a type of property that includes both unimproved land and improvements including buildings, fixtures, roads, buildings, and utility systems. Real estate comes in a variety of forms, each with a certain use and value. Land, residential, commercial, and industrial are the

key types[1]. Examples of real estate are apartment, villa, condominium (condo) etc. The profession of purchasing, selling, or renting real estate is known as real estate business[2].

1.2 Machine Learning

A branch of artificial intelligence known as machine learning is the ability of a machine to mimic intelligent human behaviour through the use of data and algorithms. For instance, Alexa uses machine learning for speech recognition. It is also employed in medical diagnosis, such as the examination of bodily fluids[3].

1.3 Machine Learning in Real Estate Business:

The main benefit of machine learning is that it is very good at analyzing massive volumes of data and discovering significant correlations. With regard to specified characteristics and variables, it can discover patterns, connections, and correlations among consumers. All in all, it can aid in forecasting expenses, pricing a location or piece of real estate, understanding consumer attitude, and predicting market trends.

1.4 Linear Regression in Machine Learning

Linear regression is one of the simplest and easiest machine learning techniques. This statistical technique is applied to predictive analysis. For continuous, real, or numeric variables like sales, salary, age, product, and price, etc., linear regression makes predictions. The linear regression algorithm, often known as linear regression, illustrates a linear relationship between a dependent variable and one or more independent variables. Since linear regression demonstrates a linear relationship, it determines how the dependent variable's value changes in proportion to the independent variable's value. The link between the variables is represented by a slanted straight line in the linear regression model. We can represent a linear regression mathematically as:

$$y = m_0 + m_1 x + \varepsilon$$

Here,

y = dependent variable / target variable

x = independent variable/predictor variable

m_0 = intercept of the line

m_1 = linear regression coefficient

ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

1.5 Types of Linear Regression

Linear regression can be classified into two types:

1.5.1 Simple Linear Regression

If only one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression is known as Simple Linear Regression.

The dependent variable for simple linear regression must have a continuous or real value. The independent variable, however, can be quantified using either continuous or categorical values.

The main goals of the simple linear regression algorithm are:

a) To create a model that shows how the two variables are related, such as how income and spending are related, how experience and pay are related, etc.

b) Predicting fresh observations, such as predicting the weather based on temperature or a company's earnings based on its annual investments.

1.5.2 Multiple Linear Regression

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Example: CO_2 emission forecasting based on engine size.

1.6 Linear Regression Line

A linear line illustrating the relationship between the dependent and independent variables is known as a regression line. A regression line can show two types of relationship:

a) **Positive Linear Relationship**:

If both the dependent and independent variable increases on the y-axis and x-axis respectively, them such a relationship is termed as a Positive linear relationship.

The line of equation will be

$$y = m_0 + m_1 x$$

b) **Negative Linear Relationship**:

If the dependent variable decreases on the y-axis and independent variable increases on the x-axis, then such a relationship is called a negative linear relationship.

The line of equation will be

$$y = m_0 - m_1 x$$

1.7 Finding the best fit line:

Finding the best fit line is crucial when using linear regression since it minimizes the difference between the predicted and actual values. The line with the least inaccuracy will have the best fit. The varied values for weights or the coefficient of lines (m_0, m_1) gives a distinct line of regression. We need to determine the best values for m_0 and m_1 to find the best fit line, thus, to do this we utilize the cost function.

1.8 Cost Function:

a) The different values for weights or coefficient of lines (m_0, m_1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.

b) Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.

c) We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.

For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

For the above linear equation, MSE can be calculated as:

$$\text{MSE} = 1\frac{1}{N}\sum_{i=1}^{n}\left(y_i - (m_1 x_i + m_0)\right)^2$$

where,

N=Total number of observations

y_i = Actual value

$(m_1 x_i + m_0)$ = Predicted value

1.9 Residuals:

The term "residual" refers to the difference between the actual value and the expected values. The residual and cost function will be high if the observed points are far from the regression line. The residual and consequently the cost function will be modest if the scatter points are near the regression line.

1.10 Gradient descent

Gradient descent is used to minimize the MSE by calculating the gradient of the cost function. A regression model uses gradient descent to update the coefficients of the line by reducing the cost function. It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

1.11 Model Performance

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called optimization. It can be achieved by below method:

1.12 R-squared Method:

2 R-squared is a statistical method that determines the goodness of fit.

3 It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.

4 The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.

5 It is also called a coefficient of determination, or coefficient of multiple determination for multiple regression.

6 It can be calculated from the below formula:

$$R^2 = \frac{\{\sum(X-\bar{x})(Y-\bar{y})\}^2}{\sum(X-\bar{x})^2 \sum(Y-\bar{y})^2}$$

1.13 Assumptions of Linear Regression

The following is a list of some important assumptions behind linear regression. There exist some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

a) Linear relationship between the features and target:

Linear regression assumes the linear relationship between the dependent and independent variables.

b) **Small or no multicollinearity between the features:**

High correlation between the independent variables is referred to as multicollinearity. Finding the genuine correlation between the predictors and the target variables may be challenging due to multicollinearity. Or we could say that it is challenging to isolate the predictor variables that are having an impact and those that are not. The model therefore makes the assumption that there is either little or no multicollinearity between the independent variables or features.

c) Homoscedasticity Assumptions:

When the error term is the same for all values of the independent variables, it is said to be homoscedastic. In a scatter plot with homoscedasticity, there shouldn't be any discernible patterns in the data distribution.

d) Normal distribution of error terms:

The error term should follow the normal distribution pattern, according to the linear regression's presumption. Confidence intervals will become either too large or too narrow if error components are not normally distributed, which could make determining coefficients challenging. The q-q plot can be used to verify it. Additionally, a straight line with no deviation in the figure indicates that the mistake is typically distributed.

e) **No autocorrelations:**

The linear regression model makes no assumptions on error term autocorrelation. The accuracy of the model will be significantly reduced if there is any correlation in the error term. Whenever there is a dependency between residual errors, autocorrelation typically happens.

2. Review of Literature

Regression analysis answers questions about the dependence of a response variable on one or more predictors, including prediction of future values of a response, discovering which predictors are important, and estimating the impact of changing a predictor or a treatment on the value of the response[4]. The real estate market is one of the most competitive in terms of pricing and the same tends to vary significantly based on a lot of factors, hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy[5]. The concept that housing prices are influenced by characteristics such as location, distance, and region is known as price prediction. Real estate price prediction is crucial for the establishment of real estate policies and can help real estate owners and agents make informative decisions[6]. The aim of this study is to employ actual transaction data and machine learning models to predict prices of real estate. The actual transaction data contain attributes and transaction prices of real estate that respectively serve as independent variables and dependent variables for machine learning models[7].

In real estate markets, appraisals of home value are essential for conducting business. To estimate the price of a property, many details about the property must be considered. Attributes like square footage or number of rooms can easily sway the price of a house up or down. The

pricing of housing properties is determined by a variety of factors. However, post-pandemic markets have experienced volatility in the Chicago suburb area, which have affected house prices greatly. In their paper "Predicting housing prices and analysing real estate markets in the Chicago suburbs using machine learning" by Kevin Xu1, Hieu Nguyen2 , analysis was done on the Naperville/Bolingbrook real estate market to predict property prices based on these housing attributes through machine learning models, and to evaluate the effectiveness of such models in a volatile market space. Gathering data from Redfin, a real estate website, sales data from 2018 up until the summer season of 2022 were collected for research. By analysing these sales in this range of time, they can also look at the state of the housing market and identify trends in price. For modelling the data, the models used were linear regression, support vector regression, decision tree regression, random forest regression, and XGBoost regression. To analyse results, comparison was made on the MAE, RMSE, and R-squared values for each model. It was found that the XGBoost model performs the best in predicting house prices despite the additional volatility sponsored by post-pandemic conditions. After modelling, Shapley Values (SHAP) were used to evaluate the weights of the variables in constructing models[8].

In the paper "Real Estate Price Prediction Using Machine Learning Algorithm**",** Mr. K. Chandra Sekhar Reddy, Uppugunduru Lokesh stated that Linear Regression (LR) and other Machine Learning algorithms are used to forecast the price of real estate. Therefore, it is possible to utilise the expected price to assist property/house sellers set their selling prices. The concept that housing prices are influenced by characteristics such as location, distance, and region is known as price prediction. Because of this, the prices they expect to pay are vastly different from what they really do. A regression model was constructed to forecast the price of residential houses. Entering the data into the website will get the desired outcome. Then, the input data will be subjected to the regression procedure. The user's input will be fed into the model, and the estimated sale price of the property will be shown on the website in a matter of seconds[9].

3. Objectives of Study:

The prime objectives of this study are to —

a. Collect the price lists of real estate business of some cities of India
b. Understand the multicollinearity and removal of such factors
c. Formulate mathematical statement of the prediction
d. Comparison of predicted prices of the cities

4. Methods and Materials:

There are various necessities of life that cannot be overlooked, including a place to live, food, and liquids. People's living standards have been rising steadily in recent years, which has raised demand for homes. The majority of people in the world buy a property, either to live in, to use as an income source, or as a location to store their money. Every year, there is an increase in housing demand, which causes property prices to grow. It can be challenging to pinpoint the precise characteristics or elements that influence a home's price due to the numerous variables that can do so, including its location and the demand for particular types of property. This makes it challenging for investors to make wise choices and for home builders to precisely determine the eventual selling price of a home. In the real estate sector, it is common practise to forecast changes in housing prices. Due to the strong association between property prices and other factors, including location, region, and population.

a. Materials

- Kaggle data of the cities of Delhi, Bangalore, Kolkata, Mumbai, Chennai and Hyderabad

The data sets after eliminating the independent variables with VIF greater than 10 are:

Table 1: Information of Delhi

Price	Area	No. of Bedrooms	Resale
10500000 6000000 15000000	1200 1000	2 3	1 0

2500000	1350	2	1
5800000	435	2	0
.	900	3	0
.	.	.	.
.	.	.	.
2500000	.	.	.
3000000	950	2	1
2600000	540	2	1
4200000	415	1	1
	900	3	1

Source: https://www.kaggle.com

Table 2: Information of Bangalore

Price	Area	No. of Bedrooms	Resale
30000000	3340	4	0
7888000	1045	2	0
4866000	1179	2	0
8358000	1675	3	0
6845000	1670	3	0
6797000	1220	2	0
20000000	2502	4	0
7105000	1438	3	0
8405000	1405	3	0
.	.	.	.
.	.	.	.
.	.	.	.
7559999	2064	3	0
5229000	1115	2	0
8545000	1010	2	0
5364000	590	1	0

Source: https://www.kaggle.com

Table 3: Information of Mumbai

Price	Area	No. of Bedrooms	Resale
4850000	720	1	1
4500000	600	1	1
6700000	650	1	1
4500000	650	1	1
5000000	665	1	1
17000000	2000	4	1
12500000	1550	3	1
10500000	1370	3	1
10500000	1356	3	1
	.	.	.
.	.	.	.
.	.	.	.
7000000	635	1	1
2485000	1180	2	0
14500000	1180	2	0
14500000	530	1	1
4100000	700	1	0
2750000	995	2	0
2750000	1020	2	0

Source: https://www.kaggle.com

Table 4: Information of Kolkata

Price	Area	No. of Bedrooms	Resale
2235000	1016	3	0
3665999	1111	2	0
3774000	1020	2	0
2524000	935	2	0
8300000	1956	3	1
3761000	1179	3	0
9727000	1107	3	0
.	.	.	.
.	.	.	.

.	.	.	.
4208000	1570	3	0
10100000	1208	3	0
6669000	815	2	0
4608000	952	2	0
9148000	1130	2	0

Source: https://www.kaggle.com

Table 5: Information of Chennai

Price	Area	No. of Bedrooms	Resale
5500000	1310	3	0
5350000	1126	2	0
8205000	1307	3	0
23400000	3600	3	0
10100000	1700	3	0
2950000	576	1	0
.	.	.	.
.	.	.	.
.	.	.	.
7834999	1599	3	0
2408000	740	2	0
5500000	1700	3	0
3400000	1599	3	0
4500000	688	2	0

Source: https://www.kaggle.com

Table 6: Information of Hyderabad

Price	Area	No. of Bedrooms	Resale
6968000	1340	2	0

29000000	3498	4	0
6590000	1318	2	0
5739000	1295	3	1
5679000	1145	2	0
6099000	1230	2	0
7000000	1350	2	0
.	.	.	.
.	.	.	.
.	.	.	.
11000000	1460	2	1
26000000	1314	2	1
13300000	2625	3	1
10800000	2050	3	0
10400000	1805	3	0

Source: https://www.kaggle.com

b. Methodology

This study is based on secondary data. These are collected from available online source. (source: https://www.kaggle.com). After completion of the data cleaning the data are processed in Jupyter platform under Python environment. The suitable methods to meet the objectives are checking and removing the influence of multicollinearity and prediction are stated below:

- Machine learning: linear regression, multiple linear regression
- Multicollinearity checking with VIF
- Multicollinearity reduction and code used in Python as follows:

```
vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape)]
return(vif)
X = Pratyakshi
calc_vif(X)
```


The independent variables having Variance Inflation Factor (VIF) greater than 10 has been eliminated.

- Codes used to find the predicted price equations:

```
from sklearn.linear_model import LinearRegression
lin = LinearRegression()
lin.fit(X_train,y_train)
lin.predict(X_test)
lin.coef_
lin.intercept
```

5. Results and Discussions

In the last section, materials collected from Kaggle for the mentioned cities have been cleaned, verified for multicollinearity. After removing the multicollinearity, the data have been analyzed with the help of linear regression of machine learning. In this chapter, various results are discussed as follows:

a. Mathematical Statements

1) Delhi:

Coefficient matrix: [26457.87373824, -7624483.15143335, -5338953.01306926]

Intercept matrix: [6767049.128597796]

Predicted Price of Delhi = 26457.87373824 times of 'Area' - 7624483.15143335 times of 'No. of Bedrooms' -5338953.01306926 times of 'Resale' + 6767049.128597796

2) Bangalore:

Coefficient matrix: [8997.99199034, -2507096.88080722, -2468660.64165462]

Intercept matrix: [3527969.2550352355]

Predicted Price of Bangalore = 8997.99199034 times of 'Area' - 2507096.88080722 times of 'No. of Bedrooms' - 2468660.64165462 times of 'Resale' + 3527969.2550352355

3) Mumbai:

Coefficient matrix: [7568.29960236, 2451761.45441781, -1450055.7795093]

Intercept matrix: [3725100.7789234295]

Predicted Price of Mumbai = 7568.29960236 times of 'Area' + 2451761.45441781 times of 'No. of Bedrooms' - 1450055.7795093 times of 'No. of Resale' +3725100.7789234295

4) Kolkata:

Coefficient matrix: [4603.41982697, -1681066.53359399, -915746.57218728]

Intercept matrix: [6996347.50125448]

Predicted Price of Kolkata = 4603.41982697 times of 'Area' - 1681066.53359399 times of 'No. of Bedrooms' -915746.57218728 times of 'Resale' + 6996347.50125448

5) Chennai:

Coefficient matrix: [9456.06312484, -1477198.33111069, 1067130.20763257]

Intercept matrix: [531032.2685513627]

Predicted Price of Chennai = 9456.06312484 times of 'Area' - 1477198.33111069 times of 'No. of Bedrooms' + 1067130.20763257 times of 'Resale' + 531032.2685513627

6) Hyderabad:

Coefficient matrix: [11321.29951152, -1544827.61648713, 792045.59627299]

Intercept matrix: [-4911097.201664787]

Predicted Price of Hyderabad = 11321.29951152 times of 'Area' -1544827.61648713 times of 'No. of Bedrooms'+ 792045.59627299 times of 'Resale' - 4911097.201664787

b. Examples of Predicted Prices

Table 7: Examples of predicted prices for some particular conditions

City	Area (Sqft)	No. of bedrooms	Resale	Price (Rs.)
Delhi	1400	3	Yes	1,55,95,669.89
	1800	2	Yes	3,38,03,302.54
	2250	5	Yes	2,28,35,896.27
	1200	3	Yes	1,03,04,095.15
	1800	3	Yes	2,61,78,819.39
Bangalore	4094	2	No	3,53,51,554.70
	4200	3	No	3,37,98,244.97
	5771	3	No	4,79,34,090.39
	3093	2	No	2,63,44,564.72
	5915	3	No	4,92,29,801.26
Mumbai	1445	3	Yes	2,05,66,522.29
	2133	1	Yes	2,08,69,989.51
	2625	2	Yes	2,70,45,354.36
	2178	2	Yes	2,36,62,324.44
	4010	1	No	3,65,25,743.64
Kolkata	900	2	No	77,77,292.28
	950	2	No	80,07,463.27
	1405	3	Yes	75,05,206.18
	1150	3	No	72,47,080.70
	1501	3	No	88,62,881.06
Chennai	1633	3	No	1,15,41,188.36
	858	2	No	56,89,937.77
	1610	3	Yes	1,23,90,829.11
	1000	1	No	85,09,897.06
	1048	2	No	74,86,589.76
Hyderabad	1013	2	No	34,67,723.97
	1650	3	No	91,34,564.14

	1107	2	No	45,31,926.12
	1575	3	No	82,85,466.68
	1010	2	No	34,33,760.07

c. Comparison:

Taking area = 1400 square ft., no. of bedrooms = 3 with resale-availability, following results have been observed:

Table 8: Comparison of predicted prices of the cities

City	Price in Rs.
Delhi	1,55,95,669.89
Bangalore	61,35,206.76
Mumbai	2,02,25,948.81
Kolkata	74,82,189.09
Chennai	1,04,05,055.86
Hyderabad	70,96,284.86

Interpretation: From the above results of table 8, it is observed that Mumbai tops the price list followed by Delhi, followed by Chennai, followed by Kolkata, followed by Hyderabad and Bangalore has the lowest price.

6. Conclusion:

The data of the cities have been collected from www.kaggle.com. After cleaning the data, multicollinearity have been checked with the help of variance inflation factor. Removing the multicollinearity, using machine learning in the Python-Jupyter platform prediction coefficient matrices along with intercept matrices have been obtained and then the prediction statements are constructed.

Predicted Price of Delhi = 26457.87373824 times of 'Area' - 7624483.15143335 times of 'No. of Bedrooms' -5338953.01306926 times of 'Resale' + 6767049.128597796

Predicted Price of Bangalore = 8997.99199034 times of 'Area' - 2507096.88080722 times of 'No. of Bedrooms' - 2468660.64165462 times of 'Resale' + 3527969.2550352355

Predicted Price of Mumbai = 7568.29960236 times of 'Area' + 2451761.45441781 times of 'No. of Bedrooms' - 1450055.7795093 times of 'No. of Resale' +3725100.7789234295

Predicted Price of Kolkata = 4603.41982697 times of 'Area' - 1681066.53359399 times of 'No. of Bedrooms' -915746.57218728 times of 'Resale' + 6996347.50125448

Predicted Price of Chennai = 9456.06312484 times of 'Area' - 1477198.33111069 times of 'No. of Bedrooms' + 1067130.20763257 times of 'Resale' + 531032.2685513627

Predicted Price of Hyderabad = 11321.29951152 times of 'Area' -1544827.61648713 times of 'No. of Bedrooms'+ 792045.59627299 times of 'Resale' - 4911097.201664787

Comparing the predicted prices of these cities it is observed that Mumbai tops the price list followed by Delhi, followed by Chennai, followed by Kolkata, followed by Hyderabad and Bangalore has the lowest price. In this study, accuracy has not been checked. This can also be checked to get more accurate results in future.

References:

[1] "Real Estate," *Corporate Finance Institute*. https://corporatefinanceinstitute.com/resources/commercial-real-estate/real-estate/ (accessed Dec. 24, 2022).

[2] "Real estate business," *Wikipedia*. Apr. 14, 2022. Accessed: Dec. 24, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Real_estate_business&oldid=1082750029

[3] "Machine Learning," *GeeksforGeeks*. https://www.geeksforgeeks.org/machine-learning/ (accessed Dec. 24, 2022).

[4] "Applied Linear Regression | Wiley Series in Probability and Statistics." https://onlinelibrary.wiley.com/doi/book/10.1002/0471704091 (accessed Dec. 24, 2022).

[5] R. Manjula, S. Jain, S. Srivastava, and P. R. Kher, "Real estate value prediction using multivariate regression models," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 263, no. 4, p. 042098, Nov. 2017, doi: 10.1088/1757-899X/263/4/042098.

[6] P.-F. Pai and W.-C. Wang, "Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices," *Appl. Sci.*, vol. 10, no. 17, Art. no. 17, Jan. 2020, doi: 10.3390/app10175832.

[7] S. Mysore, "Prediction of House Prices Using Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 6, pp. 1780–1785, Jun. 2022, doi: 10.22214/ijraset.2022.44033.

[8] K. Xu and H. Nguyen, "Predicting housing prices and analyzing real estate market in the Chicago suburbs using Machine Learning." arXiv, Oct. 12, 2022. doi: 10.48550/arXiv.2210.06261.

[9] "Real Estate Price Prediction Using Machine Learning Algorithms - Advanced Analytics and Deep Learning Models - Wiley Online Library." https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119792437.ch2 (accessed Dec. 24, 2022).

