IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

# House Price Prediction using Machine Learning by Linear Regression

**Vaibhav Badne[1], Harshal Ambalkar[2], Rutik Mahajan[3], Yukta Chandewar[4], Prof. Shraddha S. Kashid**
Students, Department of Information Technology[1,2,3,4]
Guide, Department of Information Technology [5]
Smt. Kashibai Navale College of Engineering, Pune, Maharashtra, India

**Abstract:** *Real estate is an important topic, and predictions about house prices are important. The literature tries to extract useful knowledge from historical data about property markets. Machine learning techniques are used to analyse historical property transactions in India in order to create models that can be used by buyers and sellers. The disparities in house prices between the most expensive and most affordable suburbs of Bangalore are high. The multiple linear regression that is based on mean squared error measurement is a competitive approach, as experiments have shown.*

**Keywords:** AI – Artificial Intelligence, ML – Machine Learning, ANN – Artificial Neural network, GDP – Gross Domestic Products, ICA – Independent Components Analysis.

## I. INTRODUCTION

Machine learning is a subfield of Artificial Intelligence (AI) that uses algorithms and technologies to extract useful information from data. Machine learning is a helpful tool for handling large amounts of data because it is difficult to process such a lot manually. Machine learning is a subfield of computer science that uses algorithms to solve problems. Therefore, it is based on developing algorithms that enable the machine to learn. There are two main categories of machine learning: supervised and unsupervised. The supervised learning algorithm is where the program is taught a set of predetermined rules in order to be able to predict when new data is given. Unsupervised learning is where the program tries to find relationships and patterns between data. Valuing an estate is a difficult problem for many people involved, including homeowners, buyers, agents, creditors, and investors. Is there anything else you would like to share?

### 1.1 Objectives

- To predict the price of the house.
- To predict the price using different machine learning algorithms and select the algorithm which gives the best result.
- The No Free Lunch Theorem states that algorithms will not always produce the same results when run under the same conditions. [2]
- The purpose of this study is to examine the accuracy of house price prediction utilizing several linear, Lasso, Ridge, and Random Forest regression techniques. The goal of this research is to learn more about regression approaches in machine learning [2].
- Furthermore, the available datasets should be processed to improve performance, which is performed by identifying the necessary characteristics and using one of the selection methods 2 to minimize the undesired variables, because each house has its own unique qualities that aid in estimating its price. These characteristics may or may not be shared by all houses, implying that they do not have the same impact on house pricing, resulting in erroneous results [2].

### 1.2 Aim

Living in India for so many years has given me the impression that housing and rental prices continue to increase. Since the housing crisis in 2008, housing prices have recovered remarkably well in some major housing markets. I was surprised to read in the fourth quarter of 2016 that Bombay housing prices had fallen the most in the last 4 years. Median resale prices

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

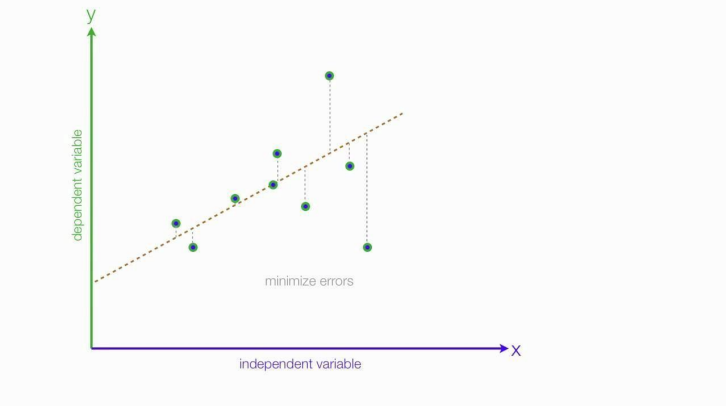
for condos and coops fell 6.3% in the third quarter of 2018, marking the first time there has been a decline since the first quarter of 2017. The decline in the US economy has been partly due to political uncertainty at home and abroad, and the 2014 election. This model allows for transparency among customers, while also making comparisons easy. If a customer finds a home price on a particular website that is higher than the price predicted by the model, they can reject the home.

## II. METHODOLOGY

Method and analysis which is performed in your research work should be written in this section. A simple strategy to follow is to use keywords from your title in the first few sentences.

### 2.1 Linear Regression

- We can summarize and investigate the relationship between two continuous quantitative variables using the simple regression statistical method.
- The predictor, explanatory, or independent variable is denoted by the letter x, while the response, outcome, or dependent variable is denoted by the letter y.

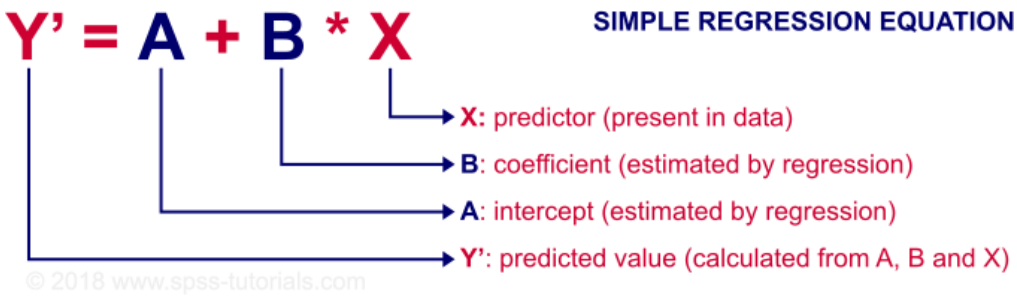



- The connection between a dependent continuous variable Y and one or more explanatory (independent) variables X is assumed to be linear in linear regression models (that is, a straight line). Rather than attempting to classify values, it is used to forecast values within a continuous range (e.g. sales, price) (e.g. cat, dog). There are two primary types of linear regression models:

### 2.2 Simple Linear Regression

- Simple linear regression employs the slope-intercept form, with a and b being the coefficients that we strive to "learn" in order to make the most accurate predictions. Our input data is represented by X, and our forecast is represented by Y.

$$Y = bX + a$$




### 2.3 Understanding The Concepts of Linear Regression

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

Linear regression is a technique for estimating linear relationships between various features and a continuous target variable. Regression means estimating a continuous real-valued output. For example, if you have data that contains the selling prices of houses in your city, you can estimate the selling price of your house based on that data and understand the market. Regression analysis is a subfield of Supervised Learning. Some of the questions that regression can answer if you are dealing with housing data are as follows:

- How much more can I sell my house for with an additional bedroom and bathroom?
- Do houses located near malls sell for more or less than others?
- What is the impact of lot size on housing prices?

Let's understand simple linear regression using an example: Suppose you have the data for employees' years of experience and their corresponding salary Tor to perform an analysis on this data, you need to make sure you collect the right data, clean the data, and transform the data. It is very important to perform exploratory data analysis to observe the distribution of the data and analyse the patterns. Fig 1 shows the plotted employee data, in which the x-axis corresponds to years of experience, and the y-axis corresponds to salary. Each point (x, y) denotes a training example. Let's consider 'it to be the total number of examples in the dataset.

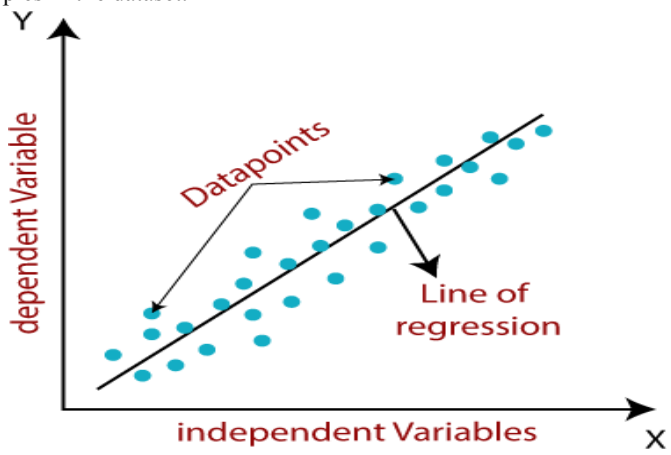



**Figure 2:** Concepts of Linear Regression

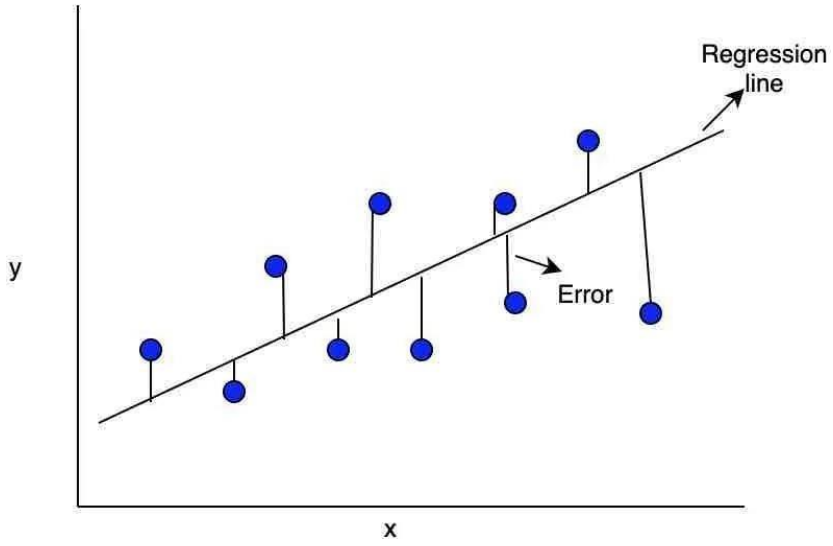


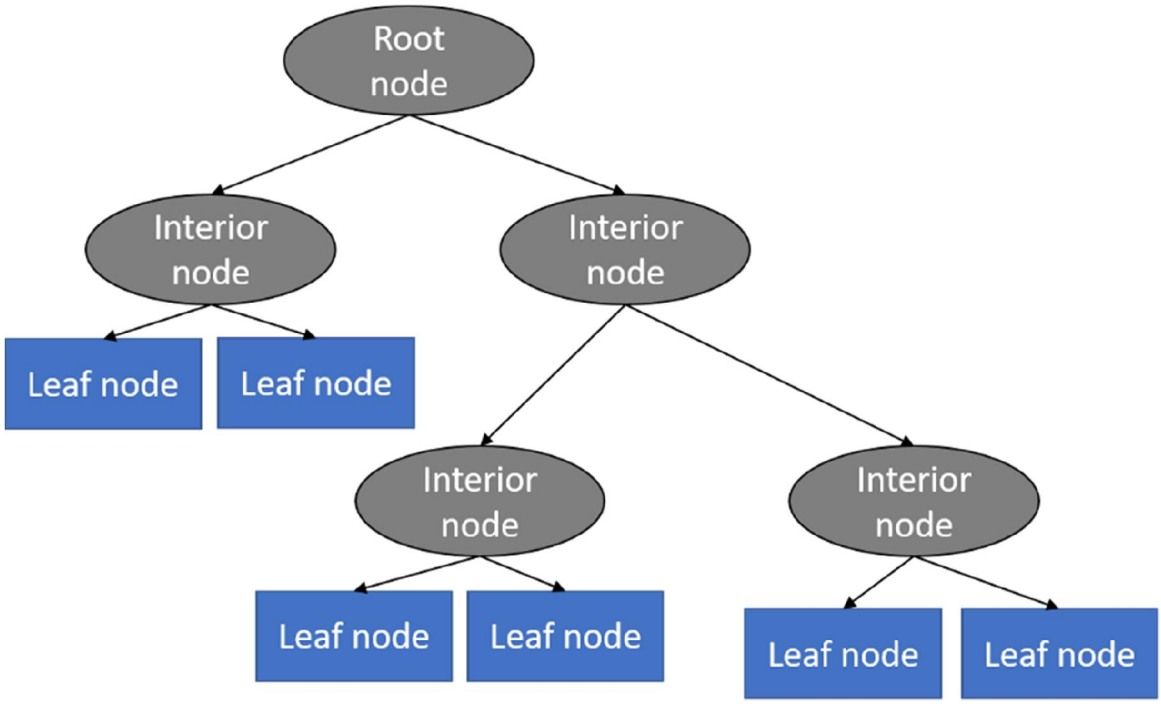

**Figure** 3: Regression line

**2.4 Decision Tree Regression**

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

One of the most widely used and useful models for supervised learning is the Decision Tree. It can be used to tackle both regression and classification problems, albeit the latter is more widely utilized.

There are three sorts of nodes in this tree-structured classifier. The Root Node is the first node in the tree that represents the full sample and can be further divided into nodes. The branches indicate decision rules, whereas the inside nodes reflect data set features. Finally, the outcome is represented by the Leaf Nodes. This algorithm is extremely beneficial for handling problems involving decisions.




A specific data point is traversed through the entire tree by answering True/False questions until it reaches the leaf node. The average of the dependent variable's value in that particular leaf node is the final forecast. The Tree is able to estimate an appropriate value for the data point after numerous iterations. In the shape of a tree structure, a decision tree constructs regression or classification models. It incrementally cuts down a dataset into smaller and smaller sections while also developing an associated decision tree. A tree with decision nodes and leaf nodes is the end result. Each branch of a decision node represents a value for the property being examined. A leaf node (for example, Hours Played) represents a numerical target decision. The root node is the topmost decision node in a tree that corresponds to the best predictor. Both category and numerical data can be handled by decision trees. Machine learning uses a variety of methods, therefore picking the right approach for the given dataset and problem is the most important thing to remember while building a machine learning model. The following are two reasons to use the Decision Tree: Decision Trees are designed to mirror human thinking abilities when making decisions, making them simple to comprehend. Because the decision tree has a tree-like form, the rationale behind it is simple to comprehend. The procedure for determining the class of a given dataset in a decision tree starts at the root node of the tree. This algorithm checks the values of the root attribute with the values of the record (actual dataset) attribute and then follows the branch and jumps to the next node based on the comparison. The algorithm compares the attribute value with the other sub-nodes and moves on to the next node. It repeats the process until it reaches the tree's leaf node. The following algorithm can help you understand the entire process:

A decision tree is a straightforward way of classifying examples. Assume that all of the input features have finite discrete domains, and that there is only one target feature termed "classification" in this section. Each element of the categorization domain is referred to as a class. Each internal (non-leaf) node of a decision tree or classification tree is labelled with an input characteristic. The arcs that emerge from a node labelled with an input feature are labelled with each of the target feature's potential values, or the arc goes to a subordinate decision node with a different input feature.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

| | location | size | total_sqft | bath | price | bhk | price_per_sqft |
|---|---|---|---|---|---|---|---|
| **5277** | Neeladri Nagar | 10 BHK | 4000.0 | 12.0 | 160.0 | 10 | 4000.000000 |
| **8483** | other | 10 BHK | 12000.0 | 12.0 | 525.0 | 10 | 4375.000000 |
| **8572** | other | 16 BHK | 10000.0 | 16.0 | 550.0 | 16 | 5500.000000 |
| **9306** | other | 11 BHK | 6000.0 | 12.0 | 150.0 | 11 | 2500.000000 |
| **9637** | other | 13 BHK | 5425.0 | 13.0 | 275.0 | 13 | 5069.124424 |

## III. Results and Discussion

### 3.1 Data Analysis and Gathering

There are various stages to my technique. The first stage is data collection, during which I gathered the data from the internet. The machine learning model will be trained using this data. The data gathered at this stage is unstructured and raw. The dataset contains 546 rows and 12 columns. The costs are given in Indian rupees, and the plot size is stated in square feet, according to the dataset. The dependent variable in the dataset is the price column, while the other columns are independent variables (also called features).

### 3.2 Data Cleaning

In order to clean the data, I looked for any missing values in any of the raw dataset's rows. However, no empty rows were identified in my dataset collection. As a result, I went on to the next phase, data pre-processing.

The process of finding and repairing corrupt or inaccurate records from a record set, table, or database is known as data cleansing, and it entails identifying incomplete, erroneous, inaccurate, or irrelevant sections of the data and then replacing, changing, or removing the filthy or coarse data.

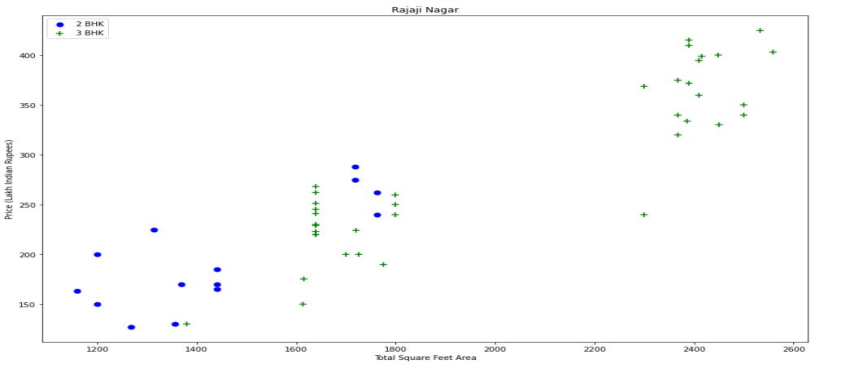



**Figure 4:** Before Data Cleaning

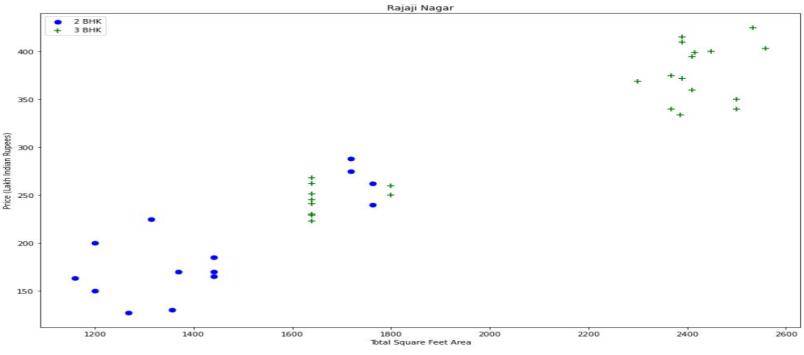
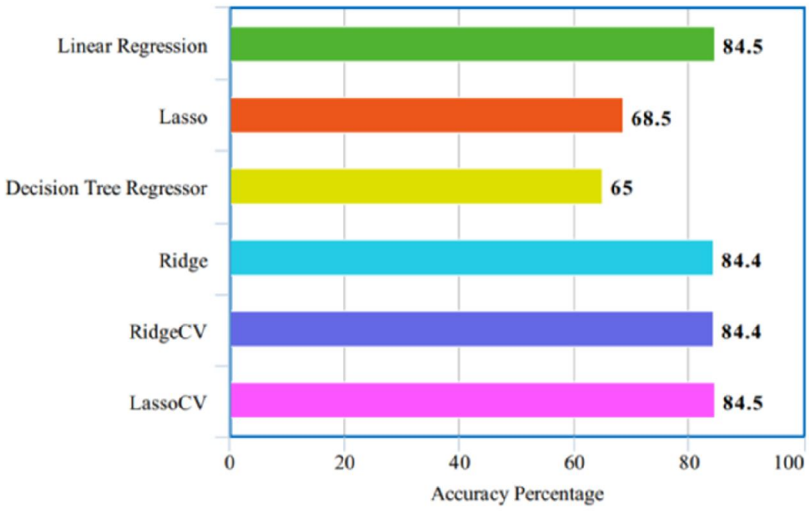



**Figure 5:** After Data Cleaning

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

## IV. RESULTS

For this paper, I have used the linear regression model to perform the prediction. However, we have many machine learning models whose accuracy can be compared to find out which one performs the best. I compared the accuracy of my model with other machine learning models like Lasso, LassoCV, Ridge, RidgeCV, and decision tree regressor. Multivariate linear regression and LassoCV perform the best with 84.5% accuracy. Thus, the model chosen for this paper (multivariate linear regression) has the highest accuracy when compared with others. The figure below shows a comparison between the linear regression model and other machine learning models.




## IV. CONCLUSION

In this project, we use machine learning algorithms to predict house prices. The study showed the different regression algorithms in use. The data needs more features to be added that are strongly correlated with house prices. The effect of crime, deposit, lending, and repo rates on house prices is weak, while the effect of inflation and year is weak and positive.

## REFERENCES

**[1].** Annina S, Mahima SD, Ramesh B. An Overview of Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering (IJESE). 2015 January.

**[2].** David HW, William GM. No Free Lunch Theorems for Optimization. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION. 1997 April.

**[3].** Sifei Lu ZLZQXYRSMG. A Hybrid Regression Technique for House Prices Prediction. In 2017 IEEE International Conference on Industrial Engineering and Engineering; 2017

**[4].** G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu. House Price Prediction Using Machine Learning (IJITEE); July 2019.

**[5].** Nils Landberg. An empirical analysis of the price development in the Swedish housing market;2015

**[6].** Suna AKKOL, Aslı AKILLI, İbrahim CEMAL.Comparison of Artificial Neural Network and Multiple Linear Regression for Prediction of Live Weight in Hair Goats;2017

**[7].** Jose Manuel Pereira, Mario Bastoa , Amelia Ferreira da Silvab .The logistic lasso and ridge regression in predicting corporate failure;2016

**[8].** Ahangar RG YMPH. The comparison of methods artificial neural network with linear regression using specific variables for prediction stock price in Tehran stock exchange. arXiv preprint arXiv:1003.1457. 2010 March.

**[9].** Annina Simon, Mahima Singh Deo, S. Venkatesan , D.R. Ramesh Babu. An Overview of Machine Learning and its Applications;2015