**REST Journal On Banking, Accounting and Business**

**Vol: 2(2), June 2023**

**REST Publisher; ISSN: 2583 4746**

**Website: https://restpublisher.com/journals/jbab/**

**DOI: https://doi.org/10.46632/jbab/2/2/8**


# House Price Prediction Using Machine Learning

***Aarti Ramawadh Mishra, Deepak Inder Moolpani**

*S. S. T. College of Arts and Commerce, Ulhasnagar, Maharashtra, India*

*Corresponding Author Email: aarti.mit21003@sstcollege.edu.in

***Abstract.*** *Prices change often in the real estate market, which makes it a dynamic business. It's one of the most important ways to use machine learning to predict the prices of real estate based on the present situation with the most accuracy. The main goal of the study paper is to predict the real prices of places and houses by using the right machine learning (ML) algorithms. The suggested article looks at some important factors and parameters for figuring out how much a property is worth. To figure out how much a house will cost, you will also need to use more regional and statistical methods. After using some machine learning methods and algorithms, the paper explains how the house pricing model works. Using the dataset from a reputable website in the proposed system makes it possible to get a thorough analysis of the data points. To improve the accuracy, algorithms like Linear regression and sklearn are used. During model building, almost all data similarities and cleaning, outlier removal and feature engineering, dimensionality reduction, gridsearchcv for hyperparameter tuning, k fold cross-validation, etc. are covered.*

***Keywords:*** *Linear regression model, Python, Machine Learning, House Price, Decision Tree, Lasso.*

## 1. INTRODUCTION

The proposed study paper talks about predictions for recent economic trends and plans. The main goal of the piece is to predict real estate prices so that the best house price prediction systems can be made using machine learning algorithms. Under the area of ML and Data Science, both the full-fledged website and the real estate price forecast are made. Based on the 2011 census, only 80% of people own their own homes. People who live in rural places own the most houses, while only about 69% of people who live in cities do. This is because the prices of homes are going up and are not clear. The main goal of designing and building this model is to make a price prediction system with an easy- to-use front end that lets users choose where they want to go and get an idea of how much it will cost. In the paper, the analysis is mostly based on data from a known website that gives a lot of sample points to help with the analysis. Before making a deal, you need to know how much the house costs. As the price of a house depends on many things, such as its size, location, number of bedrooms and bathrooms, parking space, lift, building style, balcony space, building state, price per square foot, etc. The proposed model tries to get an accurate result by taking all things into account. Machine Learning Models like support vector regression, support vector machine (SVM), logistic regression, k-means, artificial neural network, etc. can be used to predict house prices. The house-pricing plan is good for buyers, investors in real estate, and builders. This model will help all parties involved in real estate and the market figure out what the current market trends are, and which homes are good for their budgets. At first, studies focused on figuring out how the characteristics of a house affect its price based on which ML model was being used. This piece combines both figuring out how to predict house prices and the characteristics of a house. The city of Bangalore is used as a model for this paper because it is the fastest-growing city in Asia. The city's growth has already slowed its own economic growth rate, and over the last few decades, it has gone through many changes that have helped it grow, including the IT industry. Bangalore has a great social infrastructure, great places to learn, and a real infrastructure that is changing quickly. People from other states are moving to Bangalore because of these things, but the cost of living has gone up, making it hard for people to run their homes well [5]. The dataset from a reliable source that is easy to use is the first step in making a model. For our house price prediction, we picked a dataset with 13320 records of data and 9 features that we will use to train our model. There are different ways to use machine learning to predict what values will be in the future. In any

case, we need a model that can estimate the value of an item in the future with less error and more accuracy. To prepare the model for a specific goal, a large amount of data that is easy to remember is needed. Most of the time, people want to make a framework because there isn't much study on how to predict land property in India. This can estimate how much a property will cost by taking into account all of the factors that affect the goal value. Also, the accuracy of the forecast is measured by taking different error metrics into account.

## 2. LITERATURE CITED

The first thing that every average man wants and needs is real estate. Since property prices don't drop very quickly, it seems like investing in real estate is a very good idea. Investing in real estate seems like a hard task for investors when they have to choose a new house and predict the price with as little trouble as possible. There are many things that affect the price of a house, and all of them need to be taken into account in order to predict the price well. Also, making these models and using them to make predictions takes a lot of study and analysis of data. Many researchers are already working on this to get better results.

S. Rana, J. Mondal, A. Sharma, and I. Kashyap 2020 [5] have used XG Boost, Decision Tree Regression, SVR, and Random forest, among others, to predict house prices. After all of these methods have been run on the dataset, the accuracy is checked by comparing the results. From which, the decision tree algorithm had the highest accuracy at 99%, followed by the XG Boost at 63%. This was just an analysis based on trying different algorithms and models.

House Price Prediction Using Machine Learning Algorithms: A Case Study of Melbourne City, Australia, was published by T. D. Phan in 2018. This is a thorough case study of how to analyse a dataset to learn more about the housing market in Melbourne, Australia. Different regression methods have been used. Starting with reducing the amount of data and going through the steps of PCA (Principal Component Analysis) to find the best answer in the dataset. Then, for the competitive method, they used SVM (Support Vector Machine). So, how are different ways used to get the best results out of it?

M. Jain, H. Rajput, N. Garg, and P. Chawla 2020 [2] is also a method that uses some techniques to predict the price of a house. In this case, they used a simple machine learning process that included cleaning the data, showing how it looks, pre-processing it, and using k-fold cross validation to check the results. Finally, they showed a graph that shows how closely the real price and the predicted price match up. This shows that their working model is pretty accurate.

N. N. Ghosalkar and S. N. Dhage, in their paper "Real Estate Price value using Linear Regression" (2018, p. 4), use a simple method called "Linear Regression" to figure out how much a house is worth. In this paper, they tried to find the best line (relationship) between the real estate factors they took into account. They did this by using different mathematical methods, such as MSE (Mean Squared Error), RMSE (Root Mean Squared Error), etc.

After looking at a number of articles and study papers about using machine learning to predict housing prices, the focus of this article is now on how current trends in housing prices and homeownership are changing. A machine learning model is used in the suggested system to make accurate price predictions.

## 3. PROPOSED SYSTEM

The main goal of our design is to accurately predict the prices of real estate parcels in India for the next few years using different algorithms.

Linear regression is a guided learning method that predicts the value of variable (Y) based on variable (X), which is not dependent on variable (Y) [4]. It shows how the input (X) affects the output (Y) [5].

Least Absolute Shrinkage and Selection Operator (LASSO) is straight regression that takes loss into account. Loss is the point where data values get smaller and smaller until they reach a centre point, like the mean. The selection operator is an LR method that also regularises functionality. LASSO stands for least absolute shrinkage. It's the same as ridge regression, but the numbers of regularisation are different. The sum of the regression coefficients' absolute numbers is taken into account. Even the factors are set to zero to make sure there are no mistakes. So, lasso regression is used to choose which traits to use. The lariat method favours models that are simple and sparse, which means that they have fewer parameters [6, 7].

Decision Tree: It's kind of like linear regression, which is one of the ways that data mining can be used to look at many different factors. It is a tree with a root node, which is also called a decision node, and child nodes at the end, which help you make the right choice. A node with edges that go out is called a sub node. All of the other nodes that don't have any edges going out of them are called kid nodes or terminal nodes. Depending on the values of the input traits, each sub node is split into two or more sub trees [8]. Decision tree regression is a way to predict data by using a trained model in the form of a tree structure to produce sensible output and a continuous affair, which are the same thing and can't be separated [9].

First, the raw data is "feature engineered," which means that it is cleaned up and any outliers are taken out. This gets the data ready for the model to be made. Figure 1 shows that the dataset is split into two groups: training and testing. Training is 80% of the dataset, and testing is 20%. The k-fold cross- validation method, where k is 5, is used to make sure the results are correct. This means that the model is between 82% and 85% accurate. Using machine learning methods, the training set is used to make a learned model. The k-fold cross validation's hyperparameters help make decisions based on the models' best scores and parameters, which are taken into account here. The results are sent to the artefacts after the test set and the training model have been evaluated. The model is in the pickle file, and information about each column is in the json file. The back end is served by a Python Flask server that takes a set of values and gets back the values that were predicted.

**Technology used:**

Data science- The first step is data knowledge, where we take the data set and use it to do data drawing. We'll do the drawing of the data to make sure it gives accurate predictions. Machine Learning- The gutted data is put into the machine learning model, and we use methods like direct retrogression and retrogression trees to test our model. Front End (UI)- The front end of a website is mostly its structure or layout. In this to allow a piece of information that will help predict the price. It uses the information that the stoner put into the form to run the function that uses the prediction model to figure out what the house is likely to cost.

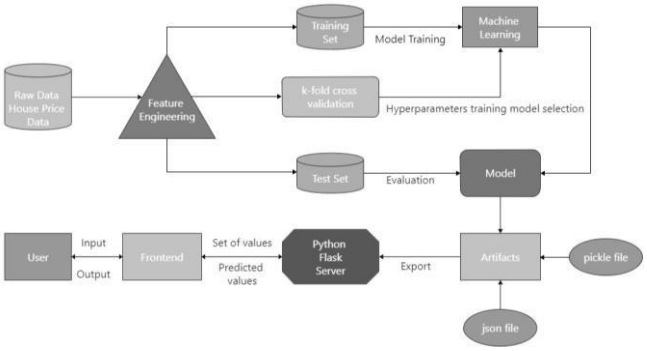


**FIGURE 1.** Architecture

## 4. DATA VISUALIZATION

As shown in Figs. 2 and 3, visualisation makes it easier to understand and use complicated data over time. Dealing with, analysing, and sending this data offers good and orderly challenges for data representation. The area of data science and the people who work in it are called "data scientists."

In Fig 2 below shows the scatterplot of price_per_sqft vs Total Square feet of the random place from the dataset Hebbal where blue dot represents 2BHK and green plus represents 3BHK. This plot is with the outliers present in the dataset.

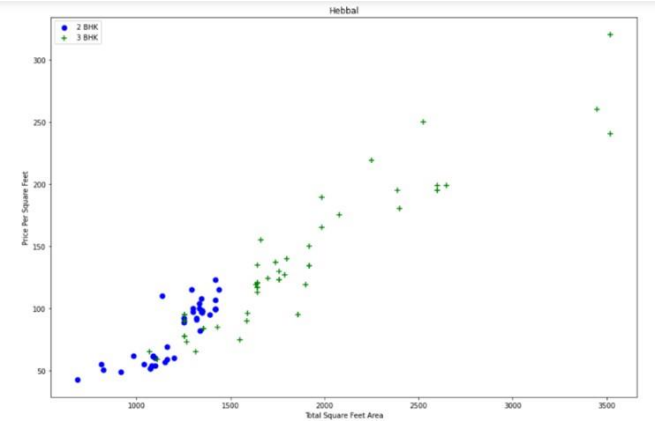


**FIGURE 2.** Price Outliers for a place (Hebbal)

Figure 3 below shows the scatterplot of price_per_sqft vs Total Square feet of a random place from the dataset Hebbal where blue dot represents 2BHK and green plus represents 3BHK. This plot is after removing the outliers present in the dataset by using the function. Also, in the above fig we can find one or two green plus which is 3BHK and still shows as outlier after the function is applied. But that is a minor difference where it has come due to the place and its area where the house is present.

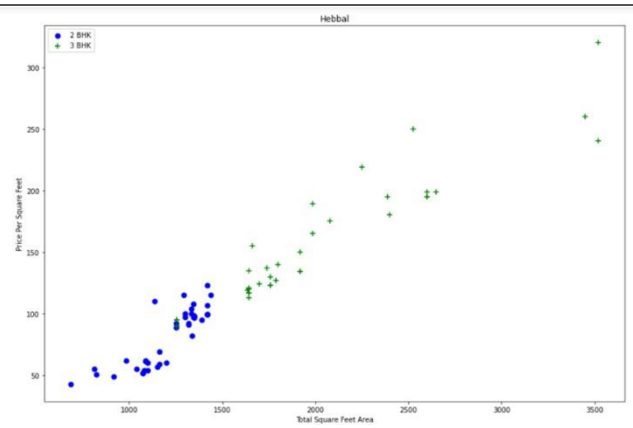


**FIGURE 3.** Price after outliers removed (Hebbal)

A correlation matrix is just a simple table that shows how the different things in the table are related to each other. The matrix shows almost all of the ways that the factors can be linked. When looking at big datasets, it is best to show a summary of the different patterns in the data. The number of the correlation matrix can be anywhere from -1 to +1. So, the positive number shows that the factors are linked in a good way, and the negative number shows that they are linked in a bad way. In Fig. 4, below, the relationships between five factors (features: total_sqft, bath, price, bhk, and price_per_sqft) are shown. For Heatmap, the matplotlib- based data visualisation is done with the sns Python tool.


**FIGURE 4.** Correlation Matrix

## 5. RESULT

Out[62]:

| | model | best_score | best_params |
|---|---|---|---|
| 0 | linear_regression | 0.847796 | {'normalize': False} |
| 1 | lasso | 0.726745 | {'alpha': 2, 'selection': 'random'} |
| 2 | decision_tree | 0.731685 | {'criterion': 'mse', 'splitter': 'random'} |

**FIGURE 5.** Comparison of the accuracy

The above Fig 5 shows the comparison between the various algorithms used to build the price prediction model, where it is found out that the Linear Regression gives the maximum accuracy of about 84.77 percent. While other algorithms Lasso and Decision Tree gives 72.26 and 73.16 percent respectively.

## 6. CONCLUSION

In this study, house prices are estimated with the help of different machine learning methods. All of the methods were explained in detail, and then the information was used as an input, and the different models were used to predict the results. Then, the way each model was presented was compared based on its features. After a proper comparison with the decision tree and Lasso regression, it was found that linear regression gave the most accurate results, with a rate of about 84 to 85%. The correlation matrix also shows how the bigger data can be seen as a small pattern. So, the model can work with enough speed to give the customer what they need.

## REFERENCES

[1]. T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, Sydney, NSW, Australia, 2018, pp. 35-42, doi: 17.1109/iCMLDE.2018.00017.

[2]. M. Jain, H. Rajput, N. Garg and P. Chawla, "Prediction of House Pricingusing Machine Learning with Python," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore,India, 2020, pp. 570-574, doi: 10.1109/ICESC48915.2020.9155839.

[3]. Nihar Bhagat, Ankit Mohokar and Shreyash Mane. House Price Forecasting using Data Mining. *International Journal of Computer Applications* 152(2):23-26, October 2016.

[4]. N. N. Ghosalkar and S. N. Dhage, "Real Estate Value Prediction Using Linear Regression," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697639.

[5]. V. S. Rana, J. Mondal, A. Sharma and I. Kashyap, "House Price Prediction Using Optimal Regression Techniques," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India, 2020, pp. 203-208, doi: 10.1109/ICACCCN51052.2020.9362864.

[6]. J. Manasa, R. Gupta and N. S. Narahari, "Machine Learning based Predicting House Prices using Regression Techniques," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 624-630, doi: 10.1109/ICIMIA48430.2020.9074952.

[7]. N. S. R H, P. R, R. R. R and M. K. P, "Price Prediction of House using KNN based Lasso and Ridge Model," *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Erode,India, 2022, pp. 1520-1527, doi:10.1109/ICSCDS53736.2022.9760832.

[8]. Z. Zhang, "Decision Trees for Objective House Price Prediction," *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, Taiyuan, China, 2021, pp. 280-283, doi: 10.1109/MLBDBI54094.2021.00059.

[9]. R. Sawant, Y. Jangid, T. Tiwari, S. Jain and A. Gupta, "Comprehensive Analysis of Housing Price Prediction in Pune Using Multi-Featured Random Forest Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697402.

[10]. C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.

 