International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

House Price Prediction Using Linear Regression Model

Jaykumar Parekh

Computer Science & Engineering , ITM (SLS) Baroda University, Paldi, Vadodara, India , 391510

Abstract

Machine learning is a subset of Artificia Intelligence. Artificial intelligence (AI) and machine learning (ML) are key technologies in solving problems and addressing a wide range of issues in many different fields. Due to its capacity to automate processes, analyze massive volumes of data, and make precise judgments. Generally, it is used in voice assistants, recommendation systems, autonomous vehicles as well as in fraud detection. It also plays a vital role in the real estate sector, accurate house price prediction helps buyers, sellers, and investors make accurate decisions. There is a need for technology to predict housing values because they rise annually. Predicting house prices can assist developers in setting a property's selling price as well as buyers in scheduling the ideal time to buy a home. Four factors influence the price of a house which are area, bedrooms, bathrooms, and location. This study uses a methodology to forecast the price of houses based on relevant features, specifically by applying a linear regression model. Through the use of machine learning methods such as Random Forest, K-Means, Decision Tree, and Linear regression. This strategy will make it easier for people to invest money in a legacy without going via a broker. The study's findings indicate that the Linear regression yields the best accuracy.

Keywords: Linear Regression, Machine Learning, Artificial Intelligence

1. INTRODUCTION

Price determination for properties was done by hand a long period ago. The issue due to this was that 20% of errors were made by hand which resulted in financial loss. Nevertheless, traditional technology has changed significantly in the present time. The technologies this day are handled using Artificial Intelligence. Machine learning a powerful branch of AI, is helping the world to grow rapidly making the task easier. Automation is a trend across all industries, though, we cannot train a model without data.

The main work of the models is to process historical data and then use them to predict future data. Our population is growing at an accelerated rate, which is driving up market demand for housing. Due to a shortage of employment, people are moving to rural areas for financial reasons. As a result, urban housing demand is rising. People who lose money because they are unaware of the house's true cost. In this project, a variety of machine learning techniques, including Linear regression, Decision Tree, K-means, and Random Forest regressions, are used to forecast the house's valuation. 20% of the data in the dataset is utilized for testing, and the remaining 80% is used for training. The following work includes data pre-processing and cleaning, feature engineering, dimensional reduction, visualization of data, splitting the dataset, and model building.

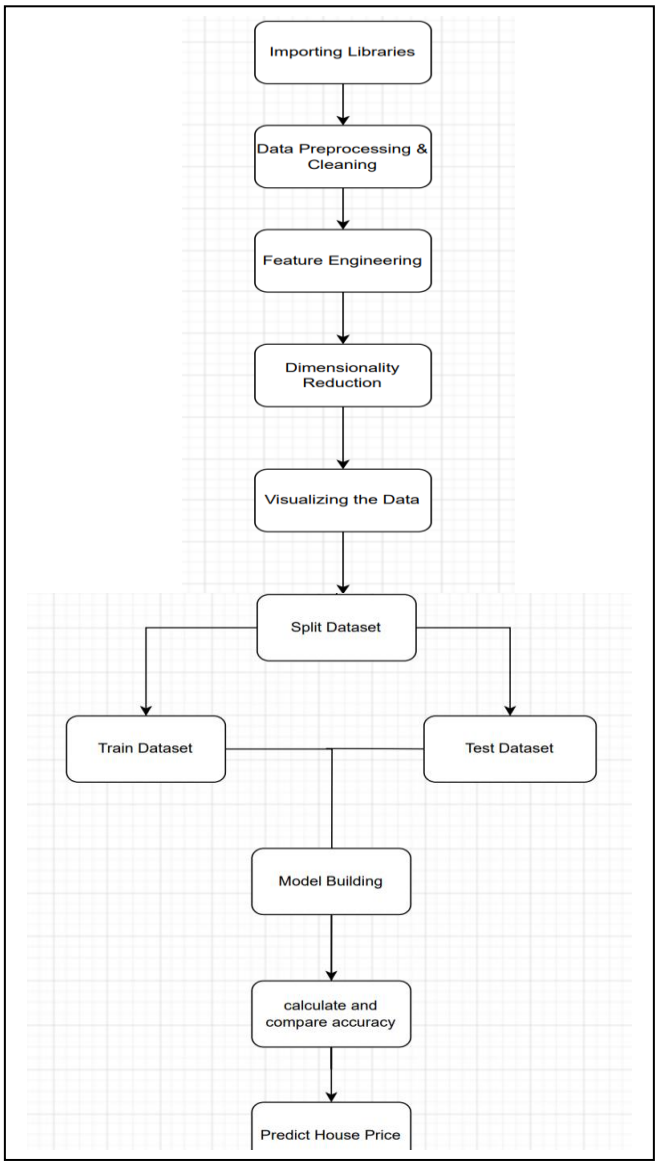

Fig.1: Flowchart of research work

2. LITERATURE SURVEY

In order to train a machine learning model more effectively, we must analyze the various machine learning algorithms. Housing cost trends provide insight into the state of the economy and are of direct relevance to both buyers and sellers. The actual cost of a home is dependent upon numerous factors. They include things like the location, the number of bathrooms, price, area type, and the number of bedrooms. Compared to cities, costs are lower in rural areas. The price of the home increases with factors like land area, shopping centers, hospitals, employment prospects, schools, etc. A few years back, real estate firms attempted to manually estimate the price of real estate. A dedicated management team is in place at the corporation to forecast the cost of any real estate property.

By examining historical data, they manually determine the price. However, that prognosis has a 30% inaccuracy rate. Hence both buyers and sellers are losing out. As a result, numerous algorithms have been created to forecast home prices. The advanced house prediction system was proposed by Thuraiya Mohd,

Suraya Masrom, and Sifei Lu. Creating a model that provides us with a reliable house price prediction based on additional features was the primary goal of this approach.

A hybrid regression technique to predict the cost of housing was proposed by Sifei Lu. The study analyses the creative feature engineering method with a limited dataset and data features. Recently, the suggested method was used as the main framework for the Kaggle challenge "House Price: Advance Regression Techniques." Predicting fair pricing for clients based on their goals and budgets is the aim of the study. This system's primary goal was to create a model that, depending on other features, would allow us to estimate housing prices accurately.

Thuraiya Mohd, Suraya Masrom, and Noraini Johari found that the Random Forest Regression and the Decision Tree Regression frequently provided the most accuracy. The Ridge and Linear Regression, with a very small reduction in Lasso, produce a comparable outcome. Regardless of strong or weak groups, there is no significant difference found across all feature selection groups. The fact that the purchase prices alone can be applied to forecast the selling prices without taking into account other factors to spread overfitting in the model is encouraging.

M Thamarai and S P Malarvizhi tried out some of the most basic machine learning algorithms, such as multiple linear regression, decision tree regression, and decision tree classifier. The machine learning tool Scikit-Learn is used to implement the work. This project assists users in forecasting the supply of homes in the city as well as the costs of those homes' dwellings.

3. DATASET

House price analysis is necessary in this growing world.
The dataset is downloaded from Kaggle, which was used in the building model and had 9 different fields in it, from which 5 fields were used which are location, size, total square foot, number of bathrooms, and price. The following data is stored in "CSV" file format. In the system, we have trained the model using different features where 80% of the data is used for training while the rest 20% is for testing purposes. Pandas, numpy, and matplotlib are the libraries used in the model, where pandas are used in analyzing and manipulating, numpy is used for working with arrays, and matplotlib is for interactive visualization. The technique used here is K-fold cross validation which trains the model and evaluates k times.

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00
5	Super built-up Area	Ready To Move	Whitefield	2 BHK	DuenaTa	1170	2.0	1.0	38.00
6	Super built-up Area	18-May	Old Airport Road	4 BHK	Jaades	2732	4.0	NaN	204.00
7	Super built-up Area	Ready To Move	Rajaji Nagar	4 BHK	Bnway G	3300	4.0	NaN	600.00
8	Super built-up Area	Ready To Move	Marathahalli	3 BHK	NaN	1310	3.0	1.0	63.25
9	Plot Area	Ready To Move	Gandhi Bazar	6 Bedroom	NaN	1020	6.0	NaN	370.00

Fig.2: Dataset

4. METHODOLOGY

A. Data pre-processing

Data cleaning

The first step is to load a dataset on which the tasks are going to be performed and analyze the following data. Now another step is to drop the unnecessary columns to get more accurate results therefore, only 5 attributes are taken into consideration. The attributes are location, size, total square feet area, number of bathrooms, and price.

	location	size	total_sqft	bath	price
0	Electronic City Phase II	2 BHK	1056	2.0	39.07
1	Chikka Tirupathi	4 Bedroom	2600	5.0	120.00
2	Uttarahalli	3 BHK	1440	2.0	62.00
3	Lingadheeranahalli	3 BHK	1521	3.0	95.00
4	Kothanur	2 BHK	1200	2.0	51.00

Fig.3: After dropping columns

The most important step to avoid the error is to drop the NA values in the data. To perform this task we will first check the number of NA values using “isnull” function. The output will show the number of NA values which is to be removed using a different function i.e. “dropna”.

Feature engineering

The column size is a mix of data considering "BHK" and “bedroom” so to make it unique we will make a new column called “BHK” which will have only numerical values in it making the data more accurate. Further modification was done in the total square feet area as the ranges were mentioned at many locations. Therefore, to convert all the ranges into a perfect area value the average of the range was calculated and replaced in place of it. Now, the column has become unique with the same type of values.

	location	size	total_sqft	bath	price	bhk
30	Yelahanka	4 BHK	2100 - 2850	4.0	186.000	4
122	Hebbal	4 BHK	3067 - 8156	4.0	477.000	4
137	8th Phase JP Nagar	2 BHK	1042 - 1105	2.0	54.005	2
165	Sarjapur	2 BHK	1145 - 1340	2.0	43.490	2
188	KR Puram	2 BHK	1015 - 1540	2.0	56.800	2
410	Kengeri	1 BHK	34.46Sq. Meter	1.0	18.500	1
549	Hennur Road	2 BHK	1195 - 1440	2.0	63.770	2
648	Arekere	9 Bedroom	4125Perch	9.0	265.000	9
661	Yelahanka	2 BHK	1120 - 1145	2.0	48.130	2
672	Bettahalsoor	4 Bedroom	3090 - 5002	4.0	445.000	4

Fig.4: Before calculating the average in total_sqft

	location	size	total_sqft	bath	price	bhk
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3
4	Kothanur	2 BHK	1200.0	2.0	51.00	2
5	Whitefield	2 BHK	1170.0	2.0	38.00	2
6	Old Airport Road	4 BHK	2732.0	4.0	204.00	4
7	Rajaji Nagar	4 BHK	3300.0	4.0	600.00	4
8	Marathahalli	3 BHK	1310.0	3.0	63.25	3
9	Gandhi Bazar	6 Bedroom	1020.0	6.0	370.00	6

Fig.5: After calculating the average in total_sqft

There are many locations where there are only a few apartments, for instance, Kanakapura Road, Giri Nagar, Nehru Nagar, etc. The locations having less than 10 data points are tagged as "other" locations. Therefore, a large number of categories can be reduced by this.

```
location
Kalkere                 10
Dairy Circle            10
Basapura                10
Nehru Nagar             10
Ganga Nagar             10
                        ..
Kanakapura  Rod          1
Kanakapura Main Road     1
1 Giri Nagar             1
Kanakapura Road,         1
whitefiled               1
Name: location, Length: 1044, dtype: int64
```

Fig.6: Location with less than 10 data points

Here we find that the minimum price per square foot is 267 rs/sqft whereas the maximum is 12000000, which shows a wide variation in property prices. We should remove outliers per location using mean and one standard deviation. To do this we took a location to check how the map of 2BHK and 3BHK looks for that particular location. The location is "Whitefield" for which a scatter plot is made indicating the points of 2 and 3 BHK. The maps help to understand the view of price and total square feet area.

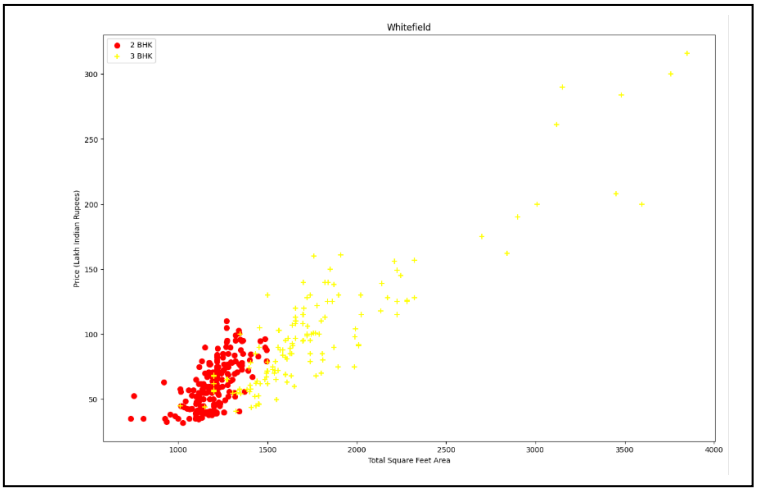

Fig.7: Plots in the Whitefield area

In Fig.7 we can see that some plots with 2BHK have more price than 3BHK which we have removed. Fig.8 shows the scatter chart after the removal of the unwanted data.

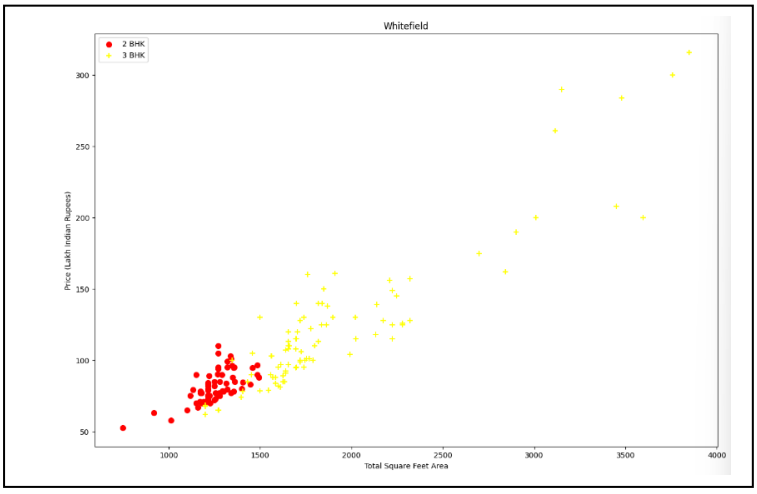

Fig.8: Scatter chart after removal of the unwanted data

B. Training the model

In model training the data is divided into 2 parts; Training and Testing. In total 80% of the data is used as training and the rest 20% for testing the model. To train the model different machine learning algorithms are used. In this model, we have used linear regression to predict the best results. In this system, k-fold cross-validation is used to check the accuracy of the Linear regression model. When there is a data shortage, resampling is done using a technique called cross-validation, or CV. The whole set of data is divided into K-folds at random, a model is fitted using (K-1) folds, the model is validated using the remaining fold, and performance is assessed using metrics. The procedure is then repeated by CV till each K-fold is employed as the testing set. The model's ultimate performance score is the mean of the K-number of scores for each metric.

The technique of fine-tuning hyperparameters to determine a model's ideal parameter values is known as grid-search. The precise values for the parameters can affect the forecast results. The grid-search method looks at every potential parameter candidate to identify the best one that will produce the best model predictions. We employed the linear regression algorithm to carry out the operation. Because property prices are continuous data we must predict them for this work. Therefore, we applied the regression procedure to continuous data. The best fit algorithm was chosen using grid search CV, which gave us the highest accuracy score in linear regression.

C. Testing

Finally, the trained model is then tested and a house price is predicted. It evaluates the effectiveness and capacity for generalization of a machine learning model. Usually, a different dataset that was not used during the model's training is used for this stage. The dataset used for testing is not the same as the training dataset a different dataset that was not used during the model's training is used for this stage. The dataset used for testing is not the same as the training dataset.

1. RESULT

There are varieties of machine learning techniques to tackle this issue. In comparison to other models, the linear regression predicts with a higher degree of accuracy.

	model	best_score	best_params
0	linear_regression	0.845056	{'fit_intercept': False}
1	lasso	0.701179	{'alpha': 1, 'selection': 'random'}
2	decision_tree	0.768908	{'criterion': 'friedman_mse', 'splitter': 'best'}

Fig.9: Best fit model accuracy

CONCLUSION

The goal of the research paper "House Price Prediction Using Linear Regression Model" is to forecast the price of a house using a variety of attributes in the data. Since linear regression has allowed us to predict a variable from an independent one, we prefer to be clear about any new information right away. In summary, the goal of this research was to support future scholars in creating a practical model that could reliably and simply forecast home prices. Our results need to be confirmed by more work on an actual model using our findings. It helps customers purchase homes within their means and minimize financial loss. We plan to incorporate other features in the future to fully estimate the price of a house.

REFERENCES

1. Arshiya Shaikh, R. Vinayaki, G. Siddhanth, Y. Phanindra Varma - " House price prediction using multivariate analysis" 2020, IJCRT.
2. Anand G. Rawool, Dattatray V. Rogye, Sainath G. Rane, Dr. Vinayk A. Bharadi - “House Price Prediction Using Machine Learning” 2021, IRE Journals.
3. Quang Truong, Minh Nguyen, Hy Dang, Bo Mei - “Housing Price Prediction via Improved Machine Learning Techniques” 2020
4. Ms. A. Vidhyavani, O. Bhargav Sathwik, Hemanth.T, Vishnu Vardhan Yadav.M - "House Price Prediction Using Machine Learning" 2021, Ijcrt.
5. Thuraiya Mohd, Suraya Masrom, Noraini Johari - “Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia” 2019, IJRTE.
6. M Thamarai, S P Malarvizhi - “House Price Prediction Modeling Using Machine Learning” 2020, DJIEEB.
7. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh - "A hybrid regression technique for house prices prediction" 2017, IEEE.
8. Sayan Putatunda – “PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market”