House Price Prediction Using Multiple Linear Regression

Anirudh Kaushal, Achyut Shankar

Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India

anirudhkushal30@gmail.com, ashankar2711@gmail.com

Abstract. There comes a point in everyone’s life when the person wishes to buy or sell a house. First consider a scenario where a person needs to buy a house. The person will look for his/her desired house for a sensible price tag. The person will have some features decided what he/she wants to have in the house. The person will be able to decide whether the type of house he/she is looking for is worth of the price or not.

Similarly, consider a scenario where a person needs to sell a house. By making use of the house price prediction system, the seller would be able to decide what all features he/she could add in the house so that the house can be sold for a higher price. Hence, from both the above scenarios we can confirm that house price prediction is useful both for the buyer and seller.

This paper will help to predict the house prices based on various parameters. The users will be able to input the type of house they desire to buy and with the help of machine learning the house price predictor will display the estimated price of the desired house.

Keywords: House Price Prediction, Machine Learning, Multiple Linear Regression

1 INTRODUCTION

Usually when people want to buy a house, they look for a house which has a reasonable cost, and which has all the desired features they want in the house. The house price prediction will help them to decide whether the house they desire to buy is worth of the price or not. Similar is the case with people who want to sell the house. By making use of the house price prediction system, the seller would be able to decide what all features he/she could add in the house so that the house can be sold for a higher price.

This paper’s objective is predicting house prices on the basis of various parameters. This will allow the buyer to get an idea of what amount of money he/she has to spend in order to buy the desired house. It will also allow the seller to get information regarding what is the house’s real worth and how he/she can maximize the profit gained by selling the house.

There are many platforms which help the buyers and sellers to predict the price of the property they are desire and the property they are looking for. Some of them are MagicBricks and 99acres. They allow the user to enter the locality of the house anywhere in India along with all the other features thus making the house price prediction system more effective.

2 LITERATURE REVIEW

Over the past few years, there have been a lot of studies conducted regarding the analysis and prediction of house prices. Wilson [7] developed an artificial neural network which helped in predicting the future trends of house prices in England. Mark and John [3] developed a regression model which was useful in analyzing house price trends of an area. Tinghao [5] predicted the real estate prices using auto regressive integrated moving average model. Zhangming [8] predicted house prices by using back propaga- tion neural network model. Sampath Kumar and Santhi [4] used multiple linear regres- sion technique to predict house price of an area, and they also predicted what would be the increase in price of the land after a period of one year. Kilpatrick [1] stated how and why the time series regression models are useful for the prediction of house prices. Wang and Tian [6] made use of the neural networks in order to find out the house price trends. Li Li and Kai-Hsuan Chuet [2] also used neural networks to predict house prices in Taipei. Instead of using normal parameters, they used economic parameters in order to make their house price prediction model.

3 METHODOLOGY AND IMPLEMENTATION

The block diagram given below is the summary of the methodology followed in the paper.

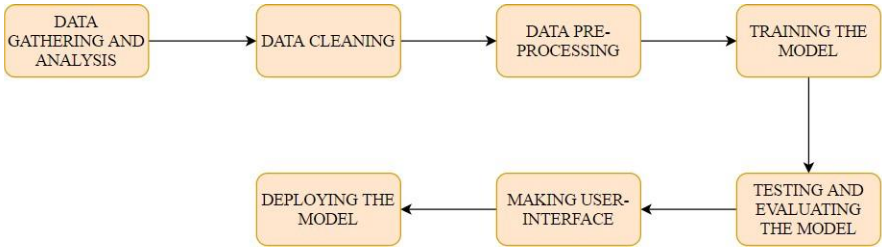
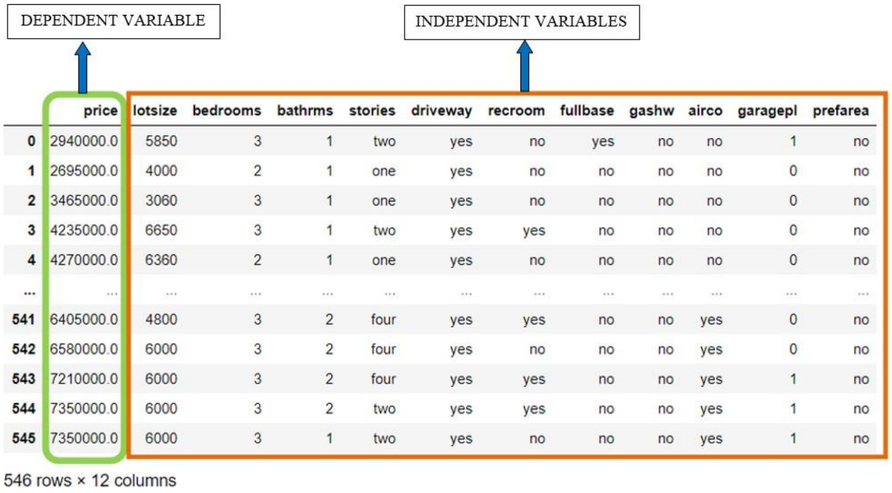


Fig. 1. Block Diagram

3.1 Data Gathering and Analysis

My procedure can be divided into several stages. The first stage is the data gathering stage in which I have collected the dataset from the internet. This will be used to train the machine learning model. The dataset collected in this stage is raw and unstructured data. There are 546 rows and 12 columns in the dataset. According to the dataset, the prices are given in Indian rupees and the plot size has been given in square feet. The price column in the dataset is the dependent variable and the rest of the columns are independent variables (also called features).



	price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase	gashw	airco	garagepl	prefarea
0	2940000.0	5850	3	1	two	yes	no	yes	no	no	1	no
1	2695000.0	4000	2	1	one	yes	no	no	no	no	0	no
2	3465000.0	3060	3	1	one	yes	no	no	no	no	0	no
3	4235000.0	6650	3	1	two	yes	yes	no	no	no	0	no
4	4270000.0	6360	2	1	one	yes	no	no	no	no	0	no
...	...	...	...	...	...	...	...	...	...	...	...	...
541	6405000.0	4800	3	2	four	yes	yes	no	no	yes	0	no
542	6580000.0	6000	3	2	four	yes	no	no	no	yes	0	no
543	7210000.0	6000	3	2	four	yes	yes	no	no	yes	1	no
544	7350000.0	6000	3	2	two	yes	yes	no	no	yes	1	no
545	7350000.0	6000	3	1	two	yes	no	no	no	yes	1	no

546 rows × 12 columns

Fig. 2. Raw and unstructured dataset

3.2 Data Cleaning

In order to clean the data, I checked if there are any missing values in any of the rows of the raw dataset. However, in my dataset set no empty rows were found. So, I moved to the next phase which is data pre-processing.

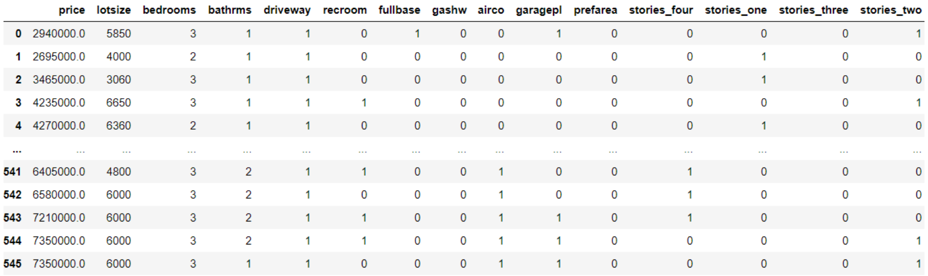
3.3 Data Pre-processing

In this phase, I converted my raw dataset into a structured form so that it is appropriate for training a machine learning model. Since I have to use multivariate regression model and it has to be trained by my dataset, it is necessary that all the independent variables are storing information in the form of numbers and not text.

However, in my dataset the columns namely driveway, recroom (recreational room), fullbase (full basement), gashw (hot water supply), airco (central air conditioning), prefarea (preferred area) have data in the form of yes or no which is text. To convert this into numerical data so that 'yes' is represented by the number 1 and 'no' is represented by the number 0, I used the 'LabelBinarizer' function available in the scikit-learn python library.

The column named stories which represents the number of floors in the house has data in the form of one, two, three and four which is text data. To convert this text data into numerical data I used the concept of 'one hot encoding'. I divided the stories column to create four new columns namely stories_one, stories_two, stories_three and stories_four. The new columns will now store the data in form of binary numbers where 0 will represent 'false' or 'no' and 1 will represent 'true' or 'yes'. After this, the original stories column is deleted because it is no longer required.

After performing all the above tasks, all the information in the dataset is in the form of numerical data and it is qualified to train the model.



	price	lotsize	bedrooms	bathrms	driveway	recroom	fullbase	gashw	airco	garagepl	prefarea	stories_four	stories_one	stories_three	stories_two
0	2940000.0	5850	3	1	1	0	1	0	0	1	0	0	0	0	1
1	2695000.0	4000	2	1	1	0	0	0	0	0	0	0	1	0	0
2	3465000.0	3060	3	1	1	0	0	0	0	0	0	0	1	0	0
3	4235000.0	6650	3	1	1	1	0	0	0	0	0	0	0	0	1
4	4270000.0	6360	2	1	1	0	0	0	0	0	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
541	6405000.0	4800	3	2	1	1	0	0	1	0	0	1	0	0	0
542	6580000.0	6000	3	2	1	0	0	0	1	0	0	1	0	0	0
543	7210000.0	6000	3	2	1	1	0	0	1	1	0	1	0	0	0
544	7350000.0	6000	3	2	1	1	0	0	1	1	0	0	0	0	1
545	7350000.0	6000	3	1	1	0	0	0	1	1	0	0	0	0	1

Fig. 3. Dataset after pre-processing

3.4 Training the regression model

For training the model, 80% of the dataset was used and for testing the model 20% of the dataset was used.

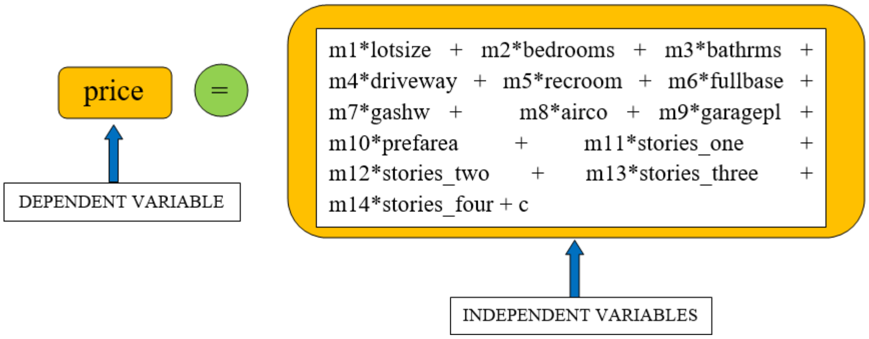

price
=
DEPENDENT VARIABLE
m1*lotsize + m2*bedrooms + m3*bathrms + m4*driveway + m5*recroom + m6*fullbase + m7*gashw + m8*airco + m9*garagepl + m10*prefarea + m11*stories_one + m12*stories_two + m13*stories_three + m14*stories_four + c
INDEPENDENT VARIABLES

Fig. 4. Mathematical equation representing multivariate regression

From the figure above, we can see that price is dependent on many factors and these factors are the independent variables/features. The model's task will be to calculate the coefficients (m1, m2, , m14) and to calculate the intercept 'c'. After calculating these, the model will be able to calculate the price for any custom input.

3.5 Evaluating the model

To evaluate how well the model is performing, I created a scatter plot which shows the comparison between the actual prices of houses mentioned in the dataset and the prices predicted by the model.

From the above figure, it can be concluded that for some datapoints, the actual price is very close to the predicted price which means that for some datapoints the model is highly accurate. However, the figure also shows that for some points the difference between the actual price and the predicted price is large which shows that for some data points the result is less accurate. Overall, we can say that the model has a decent amount of accuracy.

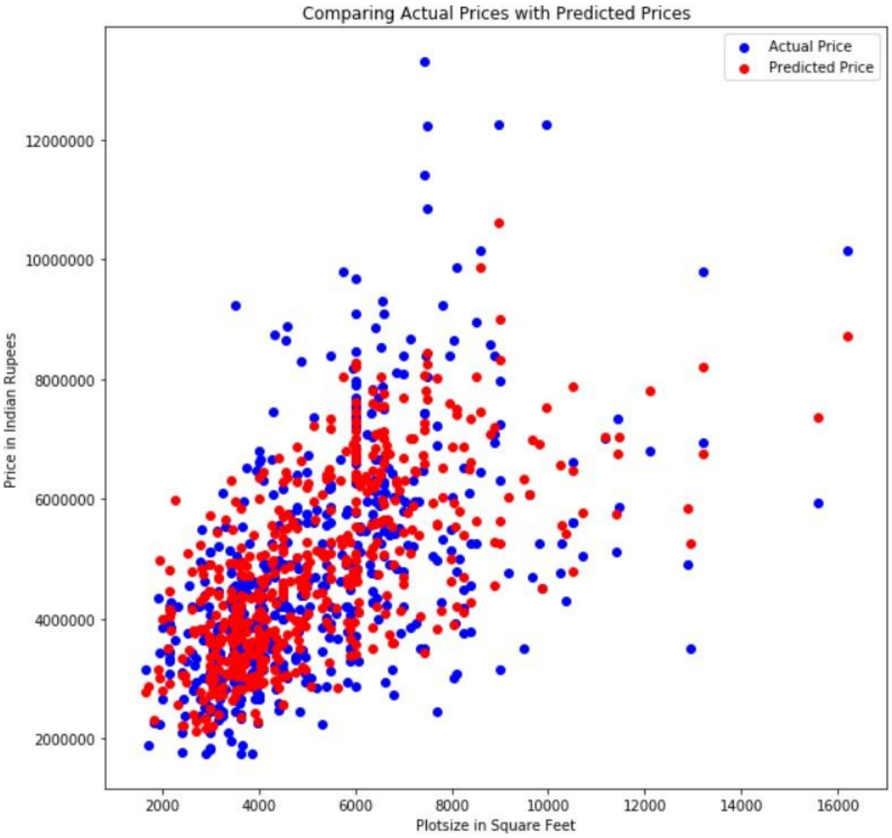


Fig. 5. Scatterplot showing difference between actual price and predicted price

4 RESULTS

For this paper, I have used the multivariate 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 performs the best with 84.5% accuracy. Thus, the model chosen for this paper (multivariate linear regression) has highest accuracy when compared with others.

The figure below shows a comparison between the multivariate linear regression model and other machine learning models.

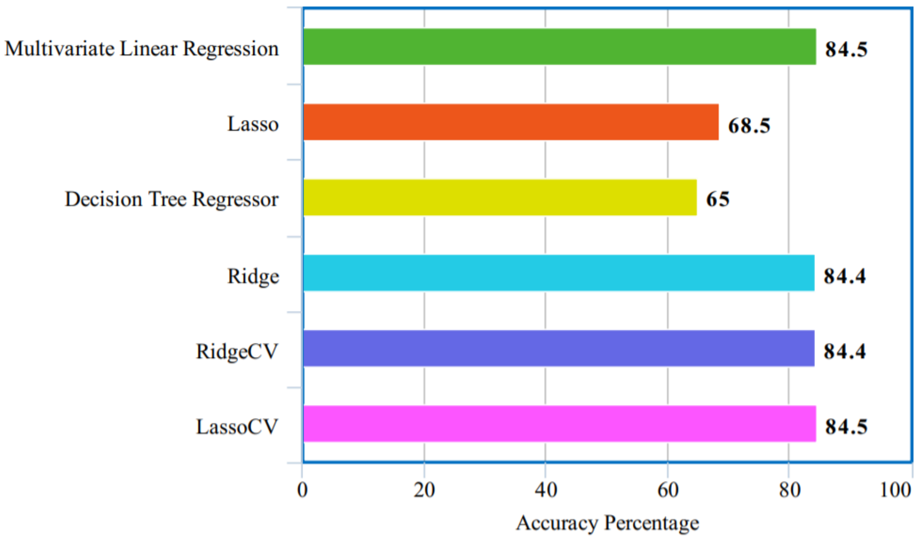
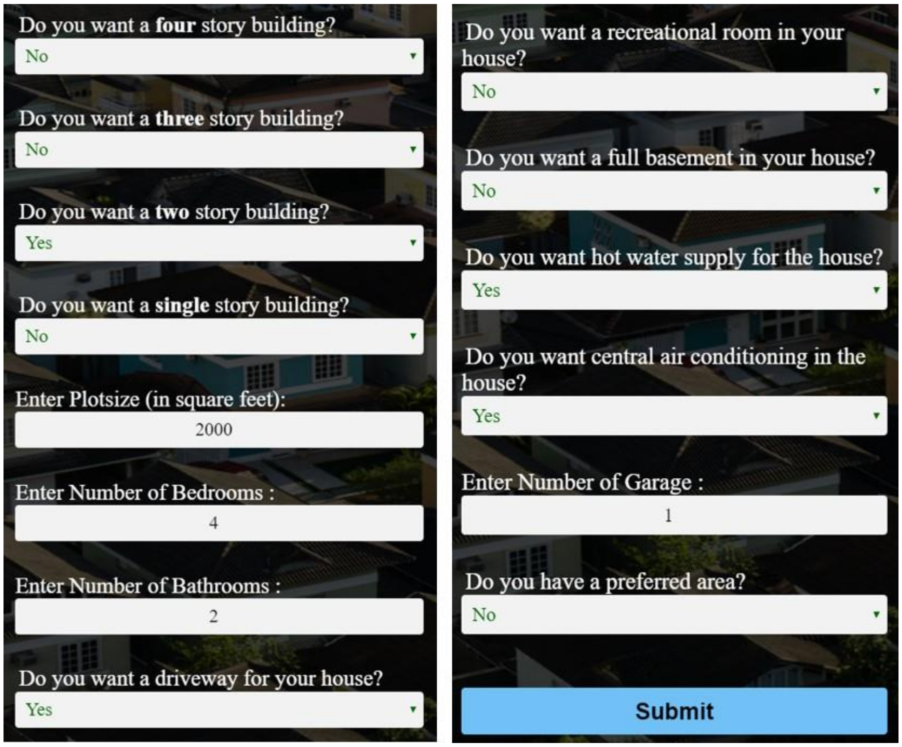


Fig. 6. Accuracy comparison among machine learning models



Do you want a **four** story building?
No

Do you want a **three** story building?
No

Do you want a **two** story building?
Yes

Do you want a **single** story building?
No

Enter Plotsize (in square feet):
2000

Enter Number of Bedrooms :
4

Enter Number of Bathrooms :
2

Do you want a driveway for your house?
Yes

Do you want a recreational room in your house?
No

Do you want a full basement in your house?
No

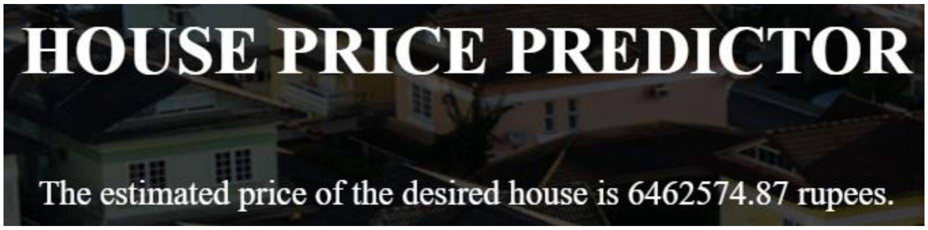
Do you want hot water supply for the house?
Yes

Do you want central air conditioning in the house?
Yes

Enter Number of Garage :
1

Do you have a preferred area?
No

Fig. 7. Input screenshot



HOUSE PRICE PREDICTOR

The estimated price of the desired house is 6462574.87 rupees.

Fig. 8. Output screenshot

5 CONCLUSION

In this paper, I have used the regression model to predict the price of different houses. It comes under the area of supervised learning which is one of the types of machine learning. All the steps required for the successful completion of the house price prediction system have been completed. It is seen that the multiple linear regression model is suitable for the purpose of predicting house prices.

5.1 Future Scope

There are some improvements and additions which can be done. The first is the ability to increase and update the dataset on a regular basis. This will make the prediction system more correct and accurate. Another improvement which can be done is add a location feature which can help the users to predict the price of houses located all over India.

6 REFERENCES

[1] Wilson, I.D., Paris, S.D, Ware, J.A., & Jenkins, D.H. Residential Property Price Time Series Forecasting With Neural Networks. Journal of Knowledge-Based Systems; 2002, 15: 335-341

[2] Mark, A.S., & John, W.B. Estimating Price Paths for Residential Real Estate. Journal of Real Estate Research; 2003: 25, 277–300.

[3] Tinghao,. Real Estate Price Index Based on ARMA Model, Statistics and Decision; 2007, 7.

[4] Zhangming, H. Research on Forecasting Real Estate Price Index Based on Neural Networks. Journal of the Graduates Sun Yat Sen University, 2006;27.

[5] Sampathkumar.V and Helen Santhi.M. Artificial Neural Network Modeling of Land Price at Sowcarpet in Chennai City, International Journal of Computer Science & Emerging Technologies; 2010, 1:44–49.

[6] Kilpatrick, J.A Factors Influencing CBD Land Prices. Journal of Real Estate; 2000, 25: 28-29.

[7] Wang, J., & Tian, P. Real Estate Price Indices Forecast by Using Wavelet Neural Network, Computer Simulation, 2005:2.

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

[9] Li Li and Kai-Hsuan Chu, “Prediction of Real Estate Price Variation Based on Economic Parameters,” Department of Financial Management, Business School, Nankai University, 2017.