International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 04 | Apr 2023 www.irjet.net p-ISSN: 2395-0072

House Price Prediction Using Machine Learning

Aditi Shahasane[1], Mayuri Gosavi[2], Ayushi Bhagat[3], Nandini Mishra[4], Amit Nerurkar[5]

[1, 2, 3, 4]Students, Dept. of Computer Engineering, Vidyalankar Institute of Technology, Wadala, Mumbai, India
[5]Professor, Dept. of Computer Engineering, Vidyalankar Institute of Technology, Wadala, Mumbai, India

---

Abstract - *Real Estate industry is dynamic in terms of the prices being fluctuated regularly. It's one of the main area to apply the machine learning concepts to predict the prices of real estate depending upon the current situations and make out maximum accuracy for the same. The research paper mainly focus on to predicting the real valued prices for the places and the houses by applying the appropriate ML algorithms. The proposed article considers some essential aspects and parameters for calculating the prices of real estate property Also some more geographical and statistical techniques will be needed to predict the price of a house. The paper consist how the house pricing model works after using some machine learning techniques and algorithms. The use of the dataset in the proposed system from the reputed website helps to get the detailed analysis of the data points. Algorithms like Linear regression and sklearn are used to effectively increase the accuracy. During model structure nearly all data similarities and cleaning, outlier removal and feature engineering, dimensionality reduction, gridsearchcv for hyperparameter tuning, k fold cross-validation, etc. are covered.*

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

1. INTRODUCTION

The proposed research paper refers to the predictions on the recent trends and for the plans of economy. The main drive behind the article is prediction of the real estate prices to build best of the house price prediction systems using the machine learning algorithms with maximum accuracy. Under the domain of ML and Data Science the designing of the real estate price prediction along with the full-fledged website is done. According to the census of 2011 only 80 percent of people own their houses. And only people based in rural areas own maximum houses but people in urban sector only about 69 % own a house. This is due to the raising prices of the properties and vague house prices. The main aim to design and develop this model is to produce price prediction system along with a user-friendly front end that will facilitate the users to choose the desired destination and get an idea about the price rates. The Analysis that has been made in the paper is mainly using the dataset from the trusted website that gives ample of sample points for better analysis. One must be aware of the exact price of house before concluding the deal. As the price of house depends on many factors like Area, location, population, size and number of bedrooms & bathrooms given, parking space, elevator, style of construction, balcony space, condition of building, price per square foot etc. The proposed model aims to create an accurate result by taking into consideration all different factors. For House price prediction one can use various prediction models (Machine Learning Models) like support vector regression, Support vector machine (SVM), Logistic regression, k-means, artificial neural network etc. House-pricing model is beneficial for the buyers, property investors, and house builders. This model will be informative and knowledgeable for the entities related to the real estate and all the stakeholders to evaluate the current market trends and budget friendly properties. Studies initially concentrated on analysis of the attributes which influence prices of the houses based on which model of ML is used and still this article brings together both predicting house price and attributes together.

For this paper, Bangalore city is taken as an example because it is Asia's fastest-growing city. The city's growth has already slowed its own economic growth rate and it has gone through various changes that have contributed to its growth over the last few decades, one of which is the IT industry. Bangalore has an excellent social infrastructure, also excellent educational institutions and a rapidly changing physical infrastructure. These factors have led to an increase in migration from other states to Bangalore, but the cost of living as increased, making it difficult for or people to manage their households effectively [5]. The model building starts with the dataset from a reliable source that is simple to use. For a dataset was chosen for our house price prediction, which contains 13320 records of data and 9 features for training our model. There are various machine learning procedures that can be used to forecast future values. In any case, it is required a model that can forecast future property estimations with greater accuracy and less error. With a specific end goal of preparing the model, a significant amount of memorable dataset is required .Generally one wants to create a framework because there is little research on forecasting land property in India. This can forecast the cost of a property by taking into account the various parameters that influence the target value. In addition, the prediction accuracy is measured by taking into account various error metrics [5].

2. LITERATURE SURVEY

Every common man's first desire and need is for real estate property. Investing in the real estate appears to be very profitable as the property rates do not fall steeply. Investing

in real estate appears to be difficult task for investors when one has to select a new house and predict the price with minimum difficulty for this there are several factors which affect the price of a house and all these factors are needed to be taken into consideration to predict the price effectively. Also building such models for prediction needs much research and data analysis as many researchers are already working on it to get the better results.

V. S. Rana, J. Mondal, A. Sharma and I. Kashyap 2020 [5] have used various regression algorithms to predict the house prices, like XG Boost, Decision Tree Regression, SVR, and Random forest. After applying all these algorithms on to the dataset a comparison for the accuracy is done at the end. From which the maximum accuracy of 99% given by the decision tree algorithm followed by the XG Boost of 63%, this was purely the experimental analysis by testing various algorithms models.

T. D. Phan, 2018 [1] is House Price Prediction using machine learning algorithms: A case study of Melbourne city, Australia. This is a through case study for analyzing the dataset to give some useful insights on to the housing industry of Melbourne city in Australia. They have used various regression models. Starting with the data reduction to applying PCA (Principal Component Analysis) steps to get the optimal solution from the dataset. Then they have applied SVM (Support Vector Machine) for the competitive approach. Thus how several methods are implemented to get the best results out of it.

M. Jain, H. Rajput, N. Garg and P. Chawla 2020 [2] is also a house price prediction system using some techniques. In this they have used the simple process of machine learning from data cleaning, visualization, pre-processing and using k-fold cross validation for the output results. Finally they have displayed the graph that shows close resemblance with actual price and the predicted price showing decent accuracy through their working model.

N. N. Ghosalkar and S. N. Dhage 2018 [4], Real Estate Price value using Linear Regression are using simple Linear Regression technique to give the price value for the houses. Through this paper they have tried to have best fitting line (relationship) between the factors of the real estate taken into consideration and used various mathematical techniques like MSE (Mean Squared Error), RMSE(Root Mean Squared Error) etc.

After reviewing various articles and research papers about machine learning for housing price prediction the article now focus is on understanding current trends in house prices and homeownership. The proposed system uses a machine learning model to predict prices with high accuracy.

3. PROPOSED SYSTEM

The main end or focus of our design is to prognosticate the accurate price of the real estate parcels present in India for the coming forthcoming times through different

Algorithms used in the model building are:

Linear Regression- It's a supervised literacy fashion and responsible for prognosticating the value of variable(Y) relying on variable(X) which is not dependent [4]. It's the relationship between the input(X) and (Y) [5].

Least Absolute Shrinkage and Selection Operator- Lasso is direct regression that considers loss. Loss is a point where data values are diminished towards a central point, like the mean. The selection operator is an LR technique that also regularizes functionality, and LASSO stands for least absolute shrinkage. It is similar to ridge regression, but it differs in the values of regularization. The absolute values of the sum of the regression coefficients are considered. It even sets the coefficients to zero to eliminate all errors. As a result, lasso regression is used to select features. The lariat procedure encourages simple, sparse models (i.e. models with smaller parameters) [6] [7].

Decision Tree- It is like linear regression, which is one of the data mining methods of analyzing multiple variables. It is a tree that consist of root node which is also called as decision node and forms a tree with leaf nodes at the end which helps to take the appropriate decision. A sub node is a node with outgoing edges. All other nodes with no outgoing edges are known as child nodes or terminal nodes. Each sub node is parted into two or more sub trees based on the values of the input attributes [8]. Decision tree regression helps to predict the data using trained model in the form of a tree structure to generate the meaningful output and continuous affair which is nothing but non separable result/affair [9].

Initially feature engineering is applied on the raw data which includes cleaning, outlier removal to make the data ready for the model building. From the fig 1, the dataset is divided into two sets i.e. training which is 80% and testing which is 20%. To find the accuracy k-fold cross validation technique is used where value of k is 5 due to which accuracy of model comes out to be around 82% to 85%. The training set is passed through machine learning algorithms to generate trained model also the hyperparameters passed by the k-fold cross validation are helpful to take decision based on best score and best parameters of the models which are considered here. After evaluating test set and trained model obtain from a training set is passed on to the artifacts where pickle file contain the model and the json file contain the column details. The back-end is supported by the python flask server which take input as set of values and provide output as predicted values.

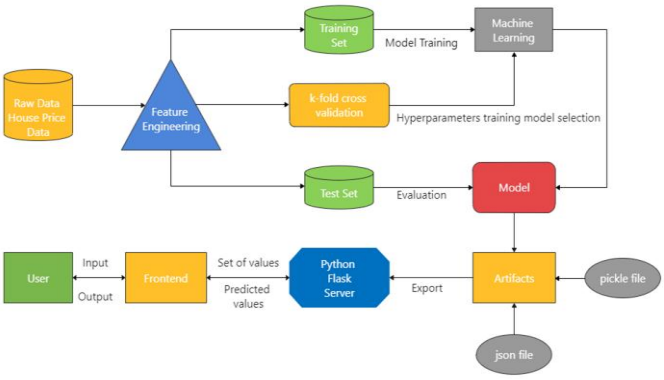

Fig-1: Architecture

Technology used-

Data Science- Data wisdom is the first stage in which we take the dataset and will do the data drawing on it. We'll do the data drawing to make sure that it provides dependable prognostications.

Machine Learning- The gutted data is fed into the machine literacy model, and we do some of the algorithms like direct retrogression, retrogression trees to test out our model.

Front End (UI) - The frontal end is principally the structure or a figure up for a website. In this to admit an information for prognosticating the price. It takes the form data entered by the stoner and executes the function which employs the prediction model to calculate the predicted price for the house.

4. DATA VISUALIZATION

Visualization gradually makes complex data more accessible, reasonable, and usable as shown in Fig 2 and Fig 3. Dealing with, analyzing, and transmitting this data presents good and orderly challenges for data representation. This test is addressed by the field of data science and experts known as 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.

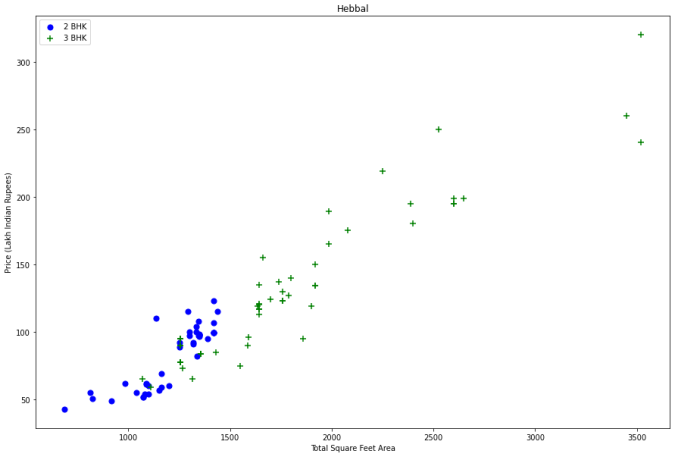

Fig-2: Price Outliers for a place (Hebbal)

In Fig 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 is has come due to the place and its area where the house is present.

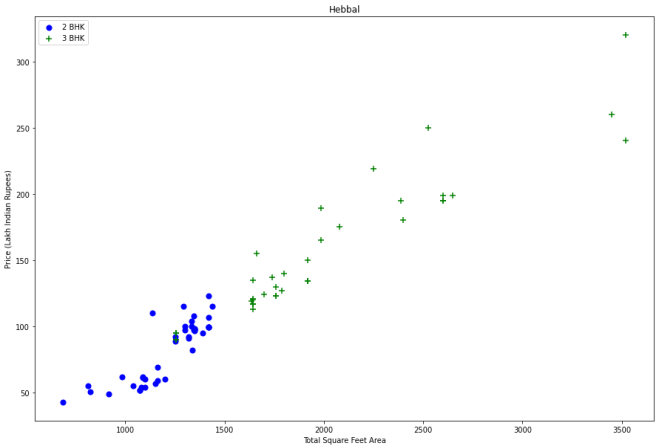

Fig-3: Price after outliers removed (Hebbal)

A correlation matrix is just a simple visual representation table that gives correlation between the different variables of the table. The matrix gives almost all the possible correlation between the variables possible. Whenever the large datasets are considered it is best option to display the summary of the different patterns of the data. The correlation matrix has the value ranging between -1 to +1. Thus the positive number shows the positive links among the variables while the negative number shows the negative link between the variables that are considered. In the Fig 4 below five variables (features- total_sqft, bath, price, bhk, and price_per_sqft) are plotted and the correlation among

them is displayed. For Heatmap the Python library sns is used for data visualization that is based on matplotlib.

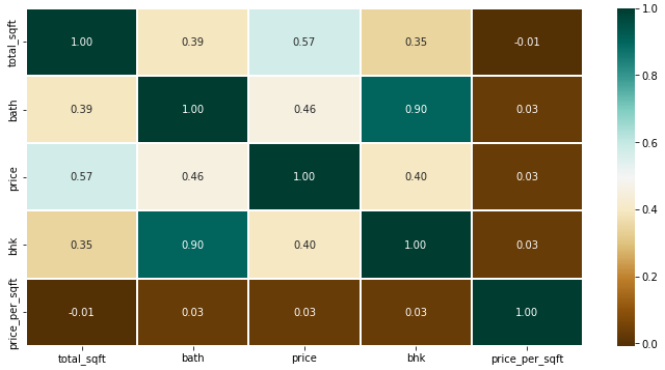

Fig-4: Correlation Matrix

5. RESULTS

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'}

Fig-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, various machine learning algorithms are used to estimate house prices. All of the methods were described in detail, and then the dataset is taken as input, applied the various models to give out the results of the prediction. The presentation of each model was then compared based on features where it is found that linear regression gives maximum accuracy of about 84 to 85% after a proper comparison with decision tree and Lasso regression. The correlation matrix also displays the visualization of the larger data into compact pattern. Thus the model can work with decent efficiency giving the required features to the customer.

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 Pricing using 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.

[11] A. P. Singh, K. Rastogi and S. Rajpoot, "House Price Prediction Using Machine Learning," 2021 3rd International Conference on Advances in Computing, Communication

Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 203-206, doi: 10.1109/ICAC3N53548.2021.9725552.

[12] A. Gupta, S. K. Dargar and A. Dargar, "House Prices Prediction Using Machine Learning Regression Models," 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2022, pp. 1-5, doi: 10.1109/ICMNWC56175.2022.10031728.

[13] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.

[14] C. Chee Kin, Z. Arabee Bin Abdul Salam and K. Batcha Nowshath, "Machine Learning based House Price Prediction Model," 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 2022, pp. 1423-1426, doi: 10.1109/ICECAA55415.2022.9936336.

[15] V. Matey, N. Chauhan, A. Mahale, V. Bhistannavar and A. Shitole, "Real Estate Price Prediction using Supervised Learning," 2022 IEEE Pune Section International Conference (PuneCon), Pune, India, 2022, pp. 1-5, doi: 10.1109/PuneCon55413.2022.10014818.

[16] S. Sharma, D. Arora, G. Shankar, P. Sharma and V. Motwani, "House Price Prediction using Machine Learning Algorithm," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 982-986, doi: 10.1109/ICCMC56507.2023.10084197.

[17] B. Almaslukh, "A Gradient Boosting Method for Effective Prediction of Housing Prices in Complex Real Estate Systems," 2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), Taipei, Taiwan, 2020, pp. 217-222, doi: 10.1109/TAAI51410.2020.00047.

[18] S. Du, Y. Gu and Y. Zhu, "Big Data Classification and Machine Learning Using Zillow Estimates," 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML), Stanford, CA, USA, 2021, pp. 254-257, doi: 10.1109/CONF-SPML54095.2021.00056.

[19] A. Verma, C. Nagar, N. Singhi, N. Dongariya and N. Sethi, "Predicting House Price in India Using Linear Regression Machine Learning Algorithms," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 917-924, doi: 10.1109/ICIEM54221.2022.9853185.

[20] X. Wang, S. Zhai and J. -L. Chen, "Research on House Price Forecast Based on Hyper Parameter Optimization Gradient Boosting Regression Model," 2020 8th International Conference on Orange Technology (ICOT), Daegu, Korea (South), 2020, pp. 1-6, doi: 10.1109/ICOT51877.2020.9468726.

[21]H.L. Harter,Method of Least Squares and some alternatives-Part II.International Static Review.1972,43(2) ,pp. 125-190.

[22] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[23] Lu. Sifei et al,A hybrid regression technique for house prices prediction. In proceedings of IEEE conference on Industrial Engineering and Engineering Management: 2017.

[24] R. Victor,Machine learning project:Predicting Boston house prices with regression in towards datascience.

[25] S. Neelam,G. Kiran,Valuation of house prices using predictive techniques, Internal Journal of Advances in Electronics and Computer Sciences:2018,vol 5,issue-6

[26] S. Abhishek.:Ridge regression vs Lasso,How these two popular ML Regression techniques work. Analytics India magazine,2018.

[27] S.Raheel.Choosing the right encoding method-Label vs One hot encoder. Towards datascience,2018.

[28] Raj, J. S., & Ananthi, J. V. (2019). Recurrent Neural Networks and Nonlinear Prediction in Support Vector Machines. Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40.

[29] Predicting house prices in Bengaluru(Machine Hackathon) https://www.machinehack.com/course/predicting-house-prices-in-bengaluru/

[30] Pow, Nissan, Emil Janulewicz, and L. Liu (2014). Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal.

[31] Wu, Jiao Yang(2017). Housing Price prediction Using Support Vector Regression.

[32] Limsombunchai, Visit. 2004.House price prediction: hedonic price model vs. artificial neural network.New Zealand Agricultural and Resource Economics Society Conference.

[33] Rochard J. Cebula (2009).The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District; The Review of Regional Studies 39.1 (2009), pp. 9– 22

[34] Gu Jirong, Zhu Mingcang, and Jiang Liuguangyan. (2011).Housing price based on genetic algorithm and support vector machine". In: Expert Systems with Applications 38 pp. 3383–3386.

[35] Danny P. H. Tay and David K. H. Ho.(1992)Artificial Intelligence and the Mass Appraisal of Residential Apartments. In: Journal of Property Valuation and Investment 10.2 pp. 525–540.

[36] Zvi Griliches, "Hedonic Price Indexes For Automobiles: an Econometric Analysis of Quality Change," National Bureau of Economic Research, vol. 0-87014-072-8, 1961.

[37] Richard C. Ready and Charles W. Abdalla, "The Amenity and Disamenity Impacts of Agriculture: Estimates from a Hedonic Pricing Model," American Journal of Agriculatural Economics, vol. 87 (2), pp. 314-326., May 2005.

[38] Ben Monty and Mark Skidmore, "Hedonic Pricing and Willingness to Pay for Bed and Breakfast Amenities in Southeast Wisconsin," Journal of Travel Research, vol. 42, no. 2, November 2003.

[39] Timothy Oladunni and Sharad Sharma, "Hedonic Housing Theory – A Machine Learning Investigation," in The 15th IEEE International Conference on Machine Learning and Applications (IEEE ICMLA'16), Los Angeles, 2016.

[40] Stacy Sirmans, David Macpherson, and Emily Zietz3, "The Composition of Hedonic Pricing Models," American Real Estate Society, vol. 13, pp. 1-44, 2005.