Real Estate Price Prediction using Supervised Learning

Vedang Matey
Student, Dept of Computer Engineering, International Institute of Information Technology
Pune, India.
vedmatey@gmail.com

Nikita Chauhan
Student, Dept of Computer Engineering, International Institute of Information Technology
Pune, India.
niks7813@gmail.com

Aditi Mahale
Student, Dept of Computer Engineering, International Institute of Information Technology
Pune, India.
aditimahale667@gmail.com

Vidya Bhistannavar
Student, Dept of Computer Engineering, International Institute of Information Technology
Pune, India.
vidyaob201@gmail.com

Dr. Ajitkumar Shitole
HOD, Dept of Computer Engineering, International Institute of Information Technology
Pune, India.
ajitkumars@isqaureit.edu.in

2022 IEEE Pune Section International Conference (PuneCon) | 978-1-6654-9897-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/PUNECON55413.2022.10014818

***Abstract*— The least transparent sector of our economy is real estate. Housing prices change daily and are occasionally inflated rather than based on an appraisal. The central focus of our approach is using fundamental factors to forecast house values. Here, we strive to establish our assessments on each essential aspect when deciding the house's price. In our project, three elements affect a house's price: its physical attributes, design, and location. There have been a lot of studies utilizing typical machine learning techniques to estimate house prices effectively. Still, they need to pay more attention to how well each model performs and ignore the less well-known but more sophisticated models. Our project involves predictions using different Regression techniques like Linear Regression, Lasso Regression, and Decision Tree. Our project includes estimating the price of houses without any expectations of market prices and cost increments. The project aims to predict residential prices for customers considering their financial plans and needs. This project means to predict house prices in Pune city with various regression techniques. The project aims to predict cogent housing prices for those who do not own homes depending on their financial capabilities and desires. Estimating pricing will be possible by examining the mentioned goods, fare ranges, and advancements. This initiative aims to enable individuals to pinpoint the specific timeline for home acquisition and sellers in assessing the cost of a home sale. Spending resources on web-based apps without consulting a broker will benefit clients.**

Additionally, it provides a brief explanation of the various graphical and numerical techniques that are required to calculate the price of a home. Our study explains the goal of machine learning, the workings of the house pricing model, and the datasets that went into developing the model we suggest. Lasso, Decision Tree, and Linear Regression were among the models looked at in the study (accuracy: 83.54 percent) (accuracy – 77.88 percent).

Keywords: machine learning, linear regression, lasso, supervised machine learning, feature extraction, decision tree.

I. INTRODUCTION

One of the three essentials of existence is the shelter. A person is protected and feels safe. Every Indian wants to own a home, but sadly for many, that goal isn't always a reality. Many residents are alarmed by the rising costs of residential properties. People spend a lot of money searching for their Dream Home. Charges have increased as a result of a loss of the proper framework, which has led to an improvement in the market's negative sentiment. This is a problem for many individuals since, if it is not resolved, many Indian citizens will no longer be able to purchase a home. In order to help potential customers, make informed choices and purchase their dream residence at the right price[17], we want to fill the void by applying system learning to anticipate future costs of residential homes. Consequently, getting rid of spiked earnings and selling an active market. In India, the real estate sector creates jobs second only to the agricultural sector. By the year 2030, India's real estate market is anticipated to reach US$ 1 trillion[1]. 13 percent of the nation's GDP will be contributed by it by 2025. (Gross home product). An encouraging sign for the quarter is rapid urbanization. It is anticipated that by 2025, there will be 525 million Indians living in urban areas. 80% of the actual property quarter is contributed by the residential segment. Residential housing is in high demand[18]. Fast urbanization, an increase in nuclear families, and easy access to credit are major factors for those. For initiatives aimed at improving townships and settlements, the government has permitted FDI up to 100%. In accordance with the "Housing for All" programme, 20 million homes would be built by the year 2022, and the GST rate would be lowered to just 5%. Tax deductions for hobbies on mortgage loans are allowed up to 2069.89 USD under the Union Budget 2021–22.

II. LITERATURE SURVEY

Nehal N Ghosalkar, et. Al. [2] explained price deviations using graphical representation between real data points and proposed best fit line. They normalized the data to get best result. This paper has used Liner Regression algorithm for prediction. They explained working of linear regression. The methodology used in this paper is interesting and they explained the relation between independent and dependent variable. They also explained RSS (Residual Sum of Squares) technique and other quality measurements. This paper got a minimum prediction[15] error using linear regression of 0.3713. They explained graphically price deviations w.r.t. (with respect to) best fit line. Rushab Sawant, et. al. [3] predicted housing price and given explanation using Decision Tree and Random forest regression. Their data consisted of 55 different features.

They selected two algorithms for training their model namely Decision Tree and Random Forest Regression. They explained the graph comparing the actual and expected prices. Their MAE (Mean Absolute Error) was 205922.62, whereas Random Forest improved it to 44031.41. Additionally, two techniques based on the MAE, Mean Squared Error (MSE), Mean Squared log Error (MSLE), and R2 Score Error are compared in this work. Random Forest outperforms the Decision Tree model, they discovered. Hedonic regression, artificial neural networks, Adaboost[4], and J48 trees—which are regarded as the finest models for prediction—were some of the techniques utilised by Aswin Ravikumar [5]. This study, according to Debanjan Banerjee and Suchibrota Dutta[6], addresses the subject of fluctuating home prices the issue of classification and uses machine learning to forecast whether residence values would increase or decrease. Mansi Jain, et al. [7] gave an outline of how to anticipate property prices using several regression techniques that incorporate increasingly complex factors for price calculation and prediction. They underlined the value of data visualisation and described how it improves data comprehension. The significance of cross validation was explained in this work. They gave an introduction to several graphical and numerical methods. This paper offers suggestions for enhancing the regression method's accuracy and precision through the use of cross validation and the stacking technique. They employed multiple regression algorithms to their dataset to provide better results, and they used a straightforward stacking approach to increase the accuracy of those algorithms. The purpose of this study, according to Ayush Varma,et.Al.[8], is to provide evaluations based on each fundamental factor taken into account while calculating the pricing. Utilized are algorithms like boosted regression and linear regression. In order to improve the estimation of home prices, Jeevan Chougale,et.Al. [9] recommended that this project take into account urban elements, such as street views and satellite image data. automated extraction of visual characteristics from photos using a deep neural network model. Ajitkumar Shitole et al.[11] prediction for the Internet of Things included real-time face recognition using sensor data. To determine the f1 score, their study uses decision trees and random forests. According to the findings, decision trees are the most accurate predictive models. Their research also clarified vector auto regression[19] for multivariate time series prediction, which has an accuracy of 83.99 percent and 88.92 percent and provides a respectable RMSE to predict temperature, humidity, etc. The current technique of calculating home values does not include the necessary forecasting of future market trends and price increases, according to Nihar Bhagat et al. [13]. Their objective was to forecast the effective home pricing for real estate buyers in consideration of their budgets. Customers were able to invest in real estate through their application without contacting an agent. Additionally, the risk associated with the transaction was reduced.

III. RESEARCH METHODOLOGY:

The methodology used to predict real estate prices is part of research methodology.

A. Architecture Diagram:

Buyers and sellers can predict a home's price by using the housing price prediction model described in this study. The CSV format is used to store the features of the data that has been acquired from diverse sources. This will enable us to employ a variety of attributes as input parameters for the convenience of purchasers or owners. The early stage, middle stage, and final stage make up the majority of an architecture design. Data cleansing and analysis are included in the early stage. The training, testing, and feature selection phases make up the intermediate stage. Visualization and final model creation are included in the last stage. Architecture is shown by Figure1.

1) Phase 1: Data Collection:

Data gathering is the methodical collection of information. This facilitates analysing data and answering questions. Social gatherings and the estimation of data on particular factors are results of the data collection procedure. Data gathering is the first and most crucial step in any machine learning project. Several websites[10], including kaggle, magicbricks, 99acres, and ready rates, a government website that provides up-to-date real estate values, are used to collect the data for this machine learning model.

2) Phase Ii: Data Cleaning And Data Loading:

Cleansing of data is process of eliminating unnecessary and faulty data present in the collected dataset. Data cleaning and elimination of such garbage values can be done using various tools. It finds out such values and replaces the messy information. This information is replaced to ensure that it is right and exact. The main motive of data cleaning is to distinguish and expel false values to build the estimation of information in dynamic way. The cleaned data then has to be stored in a new dataset. Hence, after data collection data cleaning is the important step to follow so as to get accurate results. Outlier analysis is performed for data cleaning shown by Figure 2.

3) Phase Iii: Train The Data:

Training data and testing data are created from the cleaned data at this stage. The larger dataset used to train the model's algorithms is known as the training data. The training set itself will include the goal value.

4) Phase Iv: Validation Of Model:

The process of validation involves determining whether or not the applied algorithm tests the input dataset. To achieve the highest level of accuracy is the primary motivation. Additional algorithms can be applied to the dataset after the first one to see which produces the highest accuracy. The model is presented as input-output data. The validation test compares the outputs of the system being examined to the outputs produced by the model when the model is given the identical input parameters. As a result, the outcomes were recorded. Accuracy analysis is shown by Figure 3.

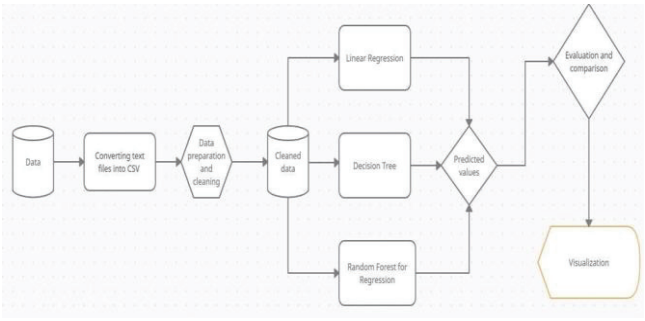

Fig. 1. System Architecture diagram

B. Algorithm Selection:

Algorithm 1: Linear Regression

1.Start.
2.Importing the dependencies like sklearn for linear regression.
3.Import linear regression from it.
4.Set the labels as price columns.
5.Again import another dependency that are needed.
6.Training and testing of data.
7.Split the data.
8.Fit the train data to the linear regression model.
9.Predict the price.
10.End.

Algorithm 2: Decision Tree:

1.Start.
2.Import libraries and dataset and use describe function to see how the data looks like.
3.Check null values in the dataset.
4.Finding unique values.
5.Dropping of the particular column value.
6.View the modified dataset.
7.Data visualization using seaborn.
8.Use heatmap to view the co relation between variables.
9.Model on the train data.
10.Calculate model score.
11.Prediction of price.
12.End.

Algorithm 3: Lasso Regression

1.Start.
2.Importing libraries & read the data.
3.Checking shape & information of the dataset.
4.Checking of null values and removing it.
5.Addressing Nan values based on data dictionary.
6.Visualize the spread target variable for sale price & function to plot scatter plot numeric variables with price.
7.Label encode ordinal features where there is order in categories.
8.Dropping columns that are not needed.
9.Converting binary variables to numeric by mapping 0 and 1.
10.Train, test and split data.
11.Plotting mean test and train scores with alpha.
12.Predict the R-squared value of test and train data.
13.Check the results.
14. End.

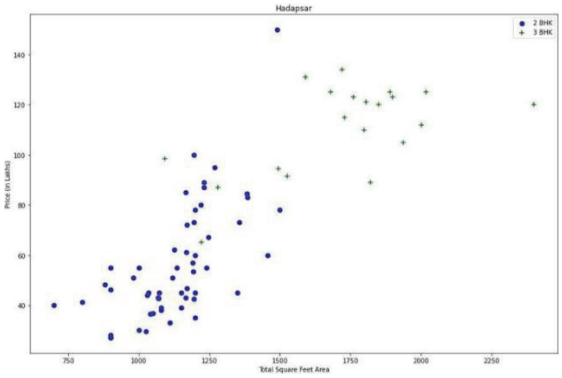

Fig. 2. Outlier analysis for particular region

TABLE I. MODEL PARAMETER ANALYSIS

S. No.	Model	Best parameters	Accuracy
1	Linear regression	{'normalize':True}	0.835453
2	Lasso	{'appha':2,'selectio n':'rando m'}	0.829241
3	Decision tree	{'criterion':'mse','s plitter':'be st'}	0.778889

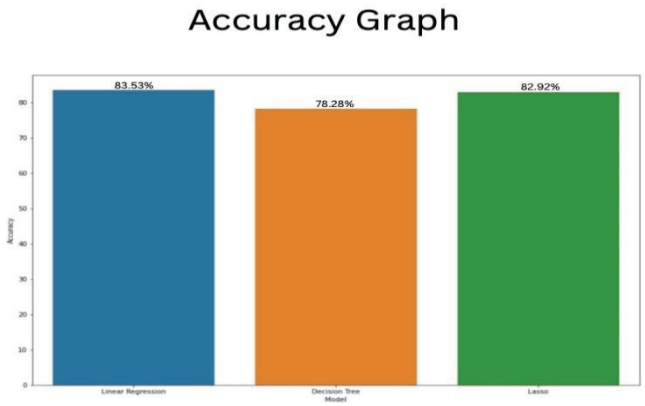

Fig. 3. Accuracy graph

IV. RESULTS AND DISCUSSIONS

Various records mining techniques are used in Python to obtain the results. Numerous factors that have an impact on home pricing are taken into account and also worked on. Machine learning is a way to finish the research work. The feature analysis process begins first. Then, to remove all of the inaccuracies from the records and make them clean, records cleaning is done. Records pre-processing is then completed. The distribution of records in unique forms is then intended to be depicted using unique plots made possible by records visualization. The final decision on the business prices [14] of the residences was made with precision. This is made feasible by the application of a simple stacking set of rules to increase the accuracy of the many regression algorithms that may be used to analyze our dataset on home price changes and produce better results. A means through which people might be helped to find houses at a price that fits within their budgets [12]. As can be seen in Figure2, 3 BHK apartments in Hadapsar cost the same as 2 BHK apartments, so it is required to remove these outliers. To do this, a scatter plot is utilized, as can be seen in Fig. 2. Grid Search is utilized in Figure 2; GridSearchCV examines every combination of the dictionary-passed data and assesses the model for each combination using the Cross Validation method [20]. The accuracy/loss for each combination of hyperparameters is obtained as a result of employing this function, and the combination with the best performance is selected. Therefore, after doing a Grid Search, we have the ideal project specifications. Model Parameter Analysis is shown by Table I.

Mean Absolute Error (MAE): Absolute error in the context of machine learning refers to the size of the discrepancy among the forecast of an observation and its actual value. The size of errors for the whole group is determined by MAE taking average the absolute errors for a set of forecasts and observations. MAE is another name for L1 loss function.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i| \qquad (1)$$

Where i = counter
N = No of non-missing data points
ŷi = actual observations time series
yi = estimated time series

Root Mean Squared Error (RMSE): Root mean square error, sometimes referred to as root mean square deviation. To calculate the root-mean-square error, calculate the difference between prediction and reality for each data point together with its norm, mean, and square root (RMSE). Each projected data point must have an actual measurement, and it employs real measurements.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}i - yi)^2}{n}} \quad (2)$$

Mean Squared Error (MSE): The Mean Squared Error (MSE) is a type of loss function that may be the simplest and most widely used. To calculate the MSE, you square the difference between your model's predictions and the actual data, then average it over the entire dataset.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2 \quad (3)$$

A statistical measure of how well a regression model fits the data is **R-squared.** The R2 statistic measures how much of the variation (which ranges from 0 to 1) is accounted for by the connection between two variables.

R-square should be set to a value of 1. The model is better suited if the r-square value is close to 1.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}i)^2}{\sum_{i=1}^{n}(yi - \bar{y}i)^2} \quad (4)$$

Model Metric analysis [16] is given by Figure4 and Table II.

TABLE II. MODEL METRICS EVALUATION

Sr no	Algo Name	MAE	RMSE	MSE x 10	R-square x 0.01
1	Linear Regression	17.953	28.751	82.6655	81.1
2	Decision Tree	17.953	28.751	82.6655	81.1
3	Lasso	17.75	28.25	78.9116	81.8

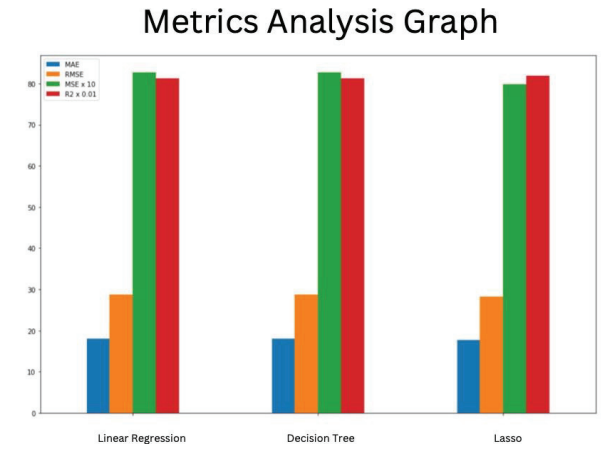

Fig. 4. Metric analysis of supervised models

V. CONCLUSION

Customer satisfaction by improving the accuracy of their decision-making and lowering the risk associated with purchasing real estate. More accurately and precisely computed sales pricing will be used. By producing precise results and reducing the chance of making the wrong investment, the system will satisfy clients. The use of machine learning in property research is still in its infancy, to sum up. We expect that our work has made a minor advancement in the field of property evaluation by offering some methodological and empirical contributions as well as by offering an alternate way for valuing housing costs. Future directions of research may take into account merging more data on real estate transactions from a bigger geographical area with more attributes or analysing other property types besides residential construction. That would make choosing homes that best fit their finances considerably simpler for the public. This research used a linear regression model, which yielded an accuracy of 83.54 percent.

VI. FUTURE SCOPE

In future, many powerful algorithms can be completed in this dataset which incorporates choice tree, Naïve Bayes, SVM etc. and find out their respective accuracies and use them to anticipate a better result. The KNN set of policies additionally may be completed to anticipate the accuracy. The k-technique set of policies additionally may be completed. With the help of these ML, the house fees are efficiently predicted. Hence, it is probably of wonderful help for the government and the people themselves. Regression algorithms are to begin with took up for our assignment but withinside the future, this will moreover, be performed using the sort algorithms. The kind algorithms can be used and it can moreover be completed to our house pricing dataset and word if they will be being completed nicely or not. The accuracy and precision of these algorithms additionally may be stepped forward consistent with our needs. Various methodologies from the world of tool getting to know are used to make our mission greater relevant. Sometimes humans moreover choose to stay near regions in which essential centers are without issue available. This is also distinctly critical difficulty to have an impact at the prices of houses and can be taken into consideration withinside the future. All the vital factors that would have an impact at the prices of houses in a selected location are almost blanketed and function worked upon them. In the future, some of the minor factors that would have an impact on house pricing on a smaller scale can be recognized and can at work upon them that how do they've an impact on house pricing and what can be completed to restriction it. In order to obtain accurate findings, additional algorithms may be deployed in future versions.

REFERENCES

[1] Indian Real Estate Industry (IBEF) January – 2021 report

[2] Nehal N Ghosalkar,Sudhir N Dhage, "Real Estate Value Prediction Using Linear Regression",International Conference on Computing Communication Control and Automation (ICCUBEA),2018.

[3] R. Sawant, T. Tiwari, A. Gupta, Y. Jangid, S. Jain, "Comprehensive Analysis of Housing Price Prediction in Pune using Multi-Featured Random Forest Approach", International Conference on Computing Communication Control and Automation (ICCUBEA) , 2018.

[4] Leo Breiman, Statistics Department, University of California, Berkeley,CA 94720, " RANDOM FORESTS ", January 2001

[5] A. Ravikumar, School of Computing National College of Ireland, Real Estate Price Prediction Using Machine Learning , December 2017.

[6] D. Banerjee, S. Dutta, "Predicting the Housing Price Direction using Machine Learning Techniques", IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI),2017.

[7] M. Jain, N. Garg, H. Rajput, P. Chawla, "Prediction of House pricing using machine learning with python", International Conference on Electronics and Sustainable Communication Systems (ICESC),2020.

[8] A. Varma, A. Sarma, S. Doshi, A. Sarma, R. Nair, "House Price Prediction Using Machine Learning And Neural Networks ", IEEE Xplore,2020.

[9] J. Chougale, N. Deshmukh, A. Shinde V. Latke, D. Sawant, " House Price Prediction using Machine learning and Image Processing", Journal of University of Shanghai for Science and Technology ,Volume 23, Issue 6, June – 2021.

[10] Pow, N. (2014). Applied Machine Learning Project 4 Prediction of real estate property prices Montréal.

[11] A. S. Shitole and Dr. M. H. Devare "Optimization of IoT Enabled Physical Location Monitoring Using DT and VAR." International Journal of Cognitive Informatics and Natural Intelligence (IJCINI) 15.4 (2021): 1-28.

[12] V. Bhistannavar, A. Mahale, N. Chauhan, V. Matey and Dr. A. Shitole. Housing Price Prediction Using Supervised Learning.

[13] N. Bhagat, S Mane and A. Mohorkar, "House Price Forecasting using Data Mining," International Journal of Computer Applications, 2016.

[14] 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.

[15] Wan Teng Lim, Yaoli Wang, Lipo Wang and Quing Chang, "Housing Price Prediction Using Neural Networks," IEEE 12th International Conference on Natural Computations, Fuzzy Systems and Knowledge Discovery, 2016.

[16] A. S. Shitole and Dr. M. H. Devare, "TPR, PPV and ROC based Performance Measurement and Optimization of Human Face Recognition of IoT Enabled Physical Location Monitoring," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 3582–3590, Jul. 30, 2019. doi: 10.35940/ijrte.b3186.078219.

[17] A. Adair, W. McGreal and J. Berry, "Hedonic modeling housing submarkets and residential valuation", Journal of Property Research, vol. 13, pp. 67-83, 1996.

[18] "Pune Population 2018" (Demographics Maps Graphs)", *World Population Review*, Jan 2017.

[19] O. Bin, "A prediction comparison of housing sales prices by parametric versus semi-parametric regressions", Journal of Housing Economics, vol. 13, pp. 68-84, 2004.

[20] T. Kauko, J. Hakfoort and P. Hooimeijer, "Capturing housing market segmentation: An alternative approach based on neural network modeling", Housing Studies, vol. 17, pp. 875-894, 2002.