International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 07 | July - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

HOUSE PRICE PREDICTION USING MACHINE LEARNING

[1] Dr. SHANKARAGOWDA B B , [2] KEERTHI B S

[1],Associate Professor and HOD,, Department of Master of Computer Application, BIET, Davangere
[2]Student, Department of MCA, BIET, Davangere

Abstract— This study shows how machine learning methods may be applied to estimate real estate or property prices using the Bangalore house price data set.This project aims to provide precise predictions about house price trends. When choosing a home, buyers take into account a number of criteria, such as location, size of the property, and proximity to places of employment, schools, parks, restaurants, and hospitals. The neighbourhood, the number of rooms a buyer wants, and the number of bathrooms he or she will require can all be used to anticipate a home's pricing in this project. The regression methods was used to complete this project because it consistently outperforms alternative models in the execution of housing cost prediction, aids in optimal model selection, and uses a flexible and probabilistic methodology. Learning Python and building expertise in data analytics, machine learning, and artificial intelligence are the main goal.

Keywords:Machine learning, house price prediction, MAPE, MAE, RMSE, Regression analysis.

I. INTRODUCTION

Real estate is a person's main ambition, but it also measures their wealth and position in moment's society. Investments in real estate constantly appear salutary since property values don't drop drastically. Changes in real estate values will have an impact on numerous people, including home buyers, bankers, policymakers, and others. Real estate investing appears to be an seductive occasion for investors. thus, vaticinating the crucial real estate price is a critical profitable index. The Asian nation, which has a aggregate of 24.67 crore homes, is ranked second in the world in terms of the number of homes, per the 2011 Census. still, real estate expenditures remain retired, as previous recessions have shown. The state's frugality has an impact on the cost of substantial estate property. There are no precise or standardized styles for determining the real estate prices of important estates, nonetheless. Read through the numerous publications and conversations on machine literacy for home price vaticination first. The design, dubbed" House Price Prediction," is grounded on machine literacy. The house- price model has numerous benefits for real estate investors, home buyers, and homebuilders. The information generated by this model is raucous, including its assessment of current asking prices for homes, which will help in setting house prices,will be of great mileage to home buyers, real estate investors, and home builders. In the suggested model,the real estate property's price is the objective characteristic, and the independent features are the number of bedrooms, space, and BHK, as well as the number of bathrooms. The programming language used for this perpetration is Python. This study builds an unprejudiced approach for prognosticating home prices using applicable technology that's both available and habituated. Understanding current developments in home power and casing costs is the study's end. The premise that real estate is a pivotal request investment is promoted in this study using a social epidemic or feedback medium. This design aims to spark Bangalore, India's house vaticination system. using a machine literacy strategy that blends analytics and data wisdom. The main pretensions of the design are to learn Python and acquire experience in AI, data analytics, and machine literacy. Due to population expansion and individuals relocating to other cities for employment opportunities, there is a constant increase in the demand for housing on the market. People are prepared to buy a new home with their budgets in place and after investigating market strategies. Therefore, the project's main objective is to predict property prices accurately while preventing any losses. There are several factors that must be considered in order to estimate home prices and try to give clients effective house pricing that respects both their preferences and their budget. As a result, a housing cost forecasting model was created for this project. Using AI, lossy regression, linear

regression,Random forest regressor and Python as machine learning tools. Customers will be able to fund bequests using this method without consulting a broker. The study's findings demonstrate that linear regression has the highest level of accuracy.

Crucial elements of this project

I. Visualization of charts that convey illuminating data.
II. In order to deliver precise findings.

II. LITERATURE SURVEY

1. To foretell home sell prices,P. Durganjali et al.suggested using classification algorithms.Data classification algorithms are used, including linear regression, decision trees, K- means.The price of a house is stated by a number of factors, including its physical attributes, geographic spot, and economic conditions. TheRMSE is used in this study to provide exact location that study better results transaction.

2. Estimated House Price J. Avanijaa et al., Gurram Sunithab et al.,K. Reddy Madhavic et al., Padmavathi Korad et al., and R. Hitesh SaiVittale et al. Suggested using the XG Boost Regression Algorithm. Numerous factors are taken into account for this blueprint, including the location, the area, and countless investments like garage space. Both buyers and marketers can gain from the suggested house foretelling setup, which helps them get the modern price for a house balance.

3. C H. Ragha Madhuri, et al., Anuradha G, et al.,and M.Vani Pujitha, et al. presented House Price Prediction Using Regression plan A respective research. In this study, foretell are made using a variety of regression plans, near as multiple direct, crest, LASSO, elastic net, rate boosting, and ada boost regression. later, Mean Square Error and Root Mean Square Error are used to gain the algorithm's accuracy number for the King County Data set.

4..Mr.P. Ravindra and associates suggested the use of sophisticated regression rules for foretell home price changes. A model was trained for this project utilizing the available data to built the most precise house price projection possible. This study used regression algorithms to estimate the costs of residences in Seattle, Washington, USA, including quickest regression, elastic net regression rules.

5.A strategy for foretelling property values using ML was put out by the authors Ayush Kumar Tiwari et al., Aman Goyal et al., Akshita Sharma et al., and Pragya Tiwari et al. The model is created using decision tree meth from the Scikit-learn module. The test set resultsare anticipated using the prediction work.

III. METHODOLOGY

This five crucial tasks makeup the usual MLworkflow:

1.Obtain Data

The data set may come from a wide range of sources, such as files, databases, sensors, and more, but it is unable to be utilized directly for the analysis process since it may contain sizable amounts of missing data, excessively high values, disordered text data, or noisy data. In order to address this issue, data Pre-processing is done.

2. Prepare, Clean, and Manipulate Data

Pre-processing is the process of cleaning up raw datawhen it has been collected in the actual world. In other words, when data are gathered from numerous sources, they are gathered in a raw format, which prevents data analysis. In order to make the data accessible and clean, particular steps are required; this stage of the process is called data transformation. Pre-processing.

3. Model Train

The content that instructs the computer way to handle data is known as the training set. The method of machine learning performs the training step using algorithms. a set of information used to fit the classifier's variables through learning.

4. Test Model

A set of unseen data that is only used to assess how well a fully specified classifier performs.

5. Model Evaluation

A critical step in the model development process is model evaluation, which helps evaluate which model best reflects our data and how well it will function going forward.

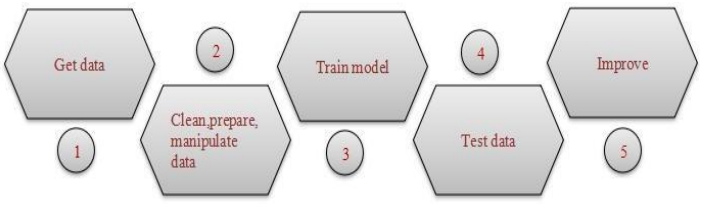


Fig 1: Shows methodology being adopted

1.Hedonic Pricing

Hedonic pricing is a method of price prediction that is based upon the hedonic theory of prices, which asserts that the worth of a property is equal to the total of all of its values for attributes. Hedonic pricing can be applied during implementation using regression modelling. Equation displays the regression model that was used to establish a price. features.

Y=a.x1 + b.x2+……..+n.xi

The features of a house include where y it is located and its estimated price.As x1,x2,,,,,,xi they relate to influencing property prices, while $a, b, \ldots\ldots n$ highlighting each variable's correlation coefficients.

A. Regression analysis

The research's prediction model was hedonic pricing.The y NJOP price, which emphasizes the cost ofconstruction of a building per square meter in the computation of building values, $x_1, x_2 \ldots . x_i$ isthe indicated dependent variable. By using the symbols, the independent variables are identified.

B. Lasso regression

L1 regularization, which facilitates feature selection and can result in sparse models, is a component of the linear regression method known as lasso regression. By encouraging less significant feature coefficients to be exactly zero, it performs feature selection by removing unimportant variables from the model.

When working with datasets with several features and in situations where features sparsity is anticipated, lasso regression is advantageous since it chooses features by setting less significant coefficients to zero.

C. Multi-layer perceptron

An artificial neural network (ANN) termed a multi-layer perceptron (MLP) is made up of numerous layers of interconnected nodes, often known as perceptron.

MLPs are capable of discovering intricate links and patterns in the data, which makes them useful for a variety of tasks like regression, classification, and pattern recognition. It have been successfully used in a variety of fields, including financial modelling, processing natural languages, and picture and audio recognition.

D.Ridge regression

In order to solve the issue of multicollinearity or strong correlation among predictor variables, the regularization method known as ridge regression is utilized in linear regression.

Numerous optimization strategies, including gradient descent and closed-form solutions, can be used to tackle the problem of ridge regression. Through methods like cross-validating, where the model's performance is assessed on a different validation set, the ideal amount of lambda can be ascertained.

In general, ridge regression is an effective method for dealing with multicollinearity and enhancing the performance and stability of regression-based models.

E.Random forest regressor

A ML method from the learning ensemble family is the Random Forest Regressor. For regression applications where the objective is to predict an ongoing numeric value instead of a categorical label, it is a strong and adaptable technique.

The idea of decision trees is the foundation of the Random Forest Regressor, which is an algorithm. Utilizing a random subset of the training information and a randomized subset of the features, it creates an ensemble with decision tree structures. This randomization aids in lowering over fitting and enhancing the model's capacity for generalization.

A.Testing Methods

Several ways, similar as Mean Absolute Chance Error(MAPE), Mean Absolute Error(MAE), and Root Mean Square Error(RMSE), will be used to test the model developed in this study. The average chance of absolute error for each expected result is known as the MAPE. MAPE can there fore-show the inflexibility of the vaticination error. A formula exists to describe MAPE.

$$MAPE = \frac{100}{n} \sum_{t}^{n} \left| \frac{x_t - y_t}{x_t} \right|$$

For each anticipated result, the average absolute error(MAE) is determined. MAE is helpful for measuring crimes in certain units. A formula can be used to determine MAE values.

$$MAE = \sum_{i=1}^{n} \left| \frac{y_i - x_i}{n} \right|$$

By taking into consideration the vaticination error of each data point, prognosticated performance is calculated using RMSE. There is an RMSE formula to be found..

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (d_i - p_i)^2}$$

PROPOSED SYSTEM APPROACH

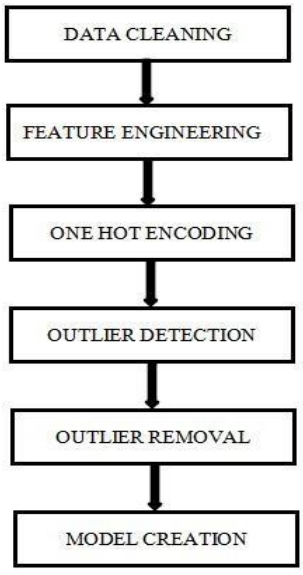
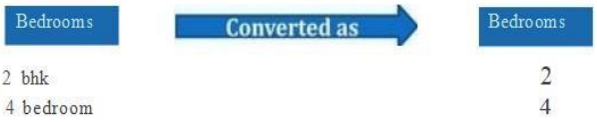


Fig 2:shows proposed system approach

DATA CLEANING

- The main ideal of data cleaning is to find and eliminate errors and indistinguishable data in order to present a dependable dataset..
- The data is gutted up using a extensively known package called pandas..
- We originally remove from our dataset the columns and attributes that aren't pivotal in calculating the final cost.The dataset excludes rows with null values in any column..
- For the columns with both characters and numeric values, only integer values are restated.
- The quality of our dataset is bettered by the operation of numerous fresh data drawing procedures.



Feature engineering

• point engineering is the process of rooting features from raw data using sphere knowledge and data mining technologies. By exercising these rates, machine learning algorithms' performance can be bettered. point engineering can be used to describe the use of machine literacy..
• In our dataset, Dimensionality reduction ways are utilized to remove the rows that aren't essential for figuring out how important a home will cost.

ONE HOT ENCODING

• Using this system, order variables are converted into numerical values..
• In our dataset, the order variable" position" is present. .
• We employed one hot garbling fashion to convert them to numerical values..

OUTLIER DETECTI ON

• Simply described, an outlier is an observation that differs from the sample's overall tendency.

We used abecedarian real estate request sphere knowledge to identify the outliers in our dataset. There are many approaches for discovering outliers, including Z- Score or Extreme Value Analysis, Probabilistic and Statistical Modelling, Information Theory Models, Standard divagation,etc.

OUTLIER REMOVAL

• After identifying the outlier, if you can, fix the errors; if not, discard the observation..
• In our dataset, we observed variations in the relationships between the values of different variables.
• To eliminate specific types of entries from the dataset.
• Scatter plots are used to find more outliers, and those are also removed from our dataset.

MODEL CREATION

• As part of the modelling process, the markers and features are separated using a machine learning fashion.
• Using the direct retrogression approach, we trained the model.
• Our model has a high delicacyrate

K-FOLD CROSS VALIDATION

• To estimate how well machines are suitable to learn new models, a statistical system known as cross-validation isused.
• It's generally used in applied machine literacy to compare and elect a model for a particular prophetic neural Network problem since it's simple to understand ,simple to apply, and constantly generates skill estimates that have lower bias than other styles.
• We find that our delicacy rate is constantly advanced percent after incorporating the k-fold cross confirmation process into our final dataset.

GRID SEARCH CROSS VALIDATION

• This approach is employed to determine the ideal parameters and the stylish modelling methodology.
• On our dataset, we combined the Lasso and linear regression, Decision Tree algorithms with the grid hunt cross confirmation methodology.
• We find that the delicacy standing for the direct regression system is above 80%.

OUTPUT

• We created a function to predict housing prices.
• Our function's syntax is: predict_ price (location, sqft, bath, bhk).
• When we pass the values into the function, it will compute the price of the house for us.

IV. RESULTS AND DISCUSSION

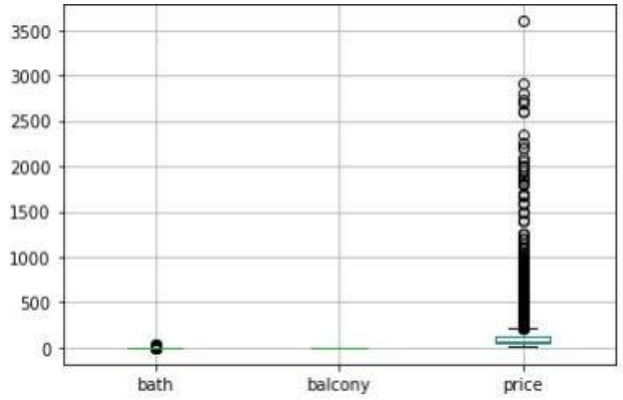


Fig 3:shows results and discussion

A data frame object from pandas library is utilized to build a box plot.A box plot is a graphical illustration that gives a summary of a dataset's distribution.

Box: The box indicates the inter quartile range (IQR), which is the range among the first quartile (Q1) and the third quartile (Q3). The median is indicated by a horizontal line within the box.

Whiskers: The whiskers extend from the box to represent the range of the data, excluding any potential outliers.where as, the whiskers extend up to 1.5 times the IQR from the edges of the box. Data points outer this range are considered potential outliers and are marked individually.

Outliers: Individual data points that fall outer the whiskers are treated as outliers and are plotted as separate points. They are typically indicated as dots on the plot.

Box plots are useful for several objectives:

Visualizing distribution: Box plots provide a quick visual overview of how the data is spread out. The length of the box indicates the spread of the central 50% of the data, while the whiskers indicates the overall range of the dataset.

Identifying skewness: By observing at the boxplot, we can observe if the data is symmetrically distributed or skewed to one side. If one whisker is longer than the other, it represents asymmetry in the data.

Detecting outliers: Boxplots helps to identify potential outliers in the dataset. Any data points that fall outer the whiskers can be considered as potential outliers.

Comparing distributions: Boxplots allow for simple comparison of distributions between various groups or categories. various boxplots can be placed side by side tocompare their medians, quartiles, and ranges.

Overall, boxplots gives a concise summary of the distribution and key statistical measures of a dataset, making them a expensive tool for exploratory data analysis and data visualization.

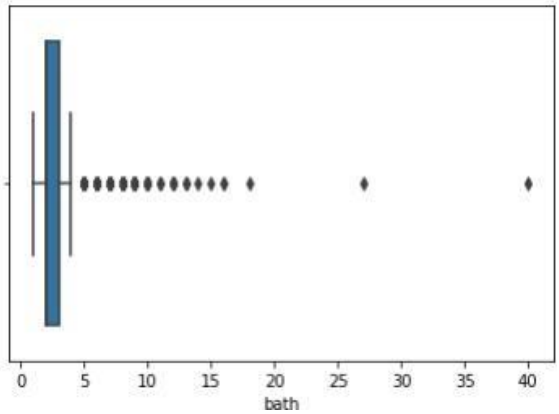


Fig 4:shows boxplot gives a concise summary

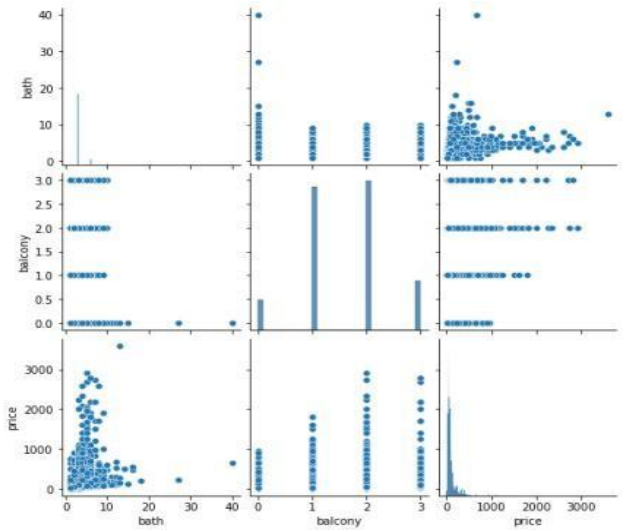


Fig 5:shows scatter diagram

We noticed at pairwise scatter plots, which let us notice the pair-wise correlations and intervals between the different variables, as shown in Figure. The scatter diagram helps us decide the data points dispersion. Obtaining a rapid understanding of the distribution of the data and whether or not outliers are present is favorable. Additionally, we can inferfrom the histogram that the price parameters—for which we should point out that we did a log transformation for the original price parameters —seems to be regularly distributed but actually holds a number of outliers.

Fig 6:heat map is utilized in order to view the correlation between variables

	Model	Best score
0	Linear_regression	0.847796
1	Lasso	0.726748
2	Decision_tree	0.713610

Outcome of the project

Figure 2: User Interface

Figure 3: Estimate Price

Fig 7: shows outcome of the project

The Python tkinter library is used to originate the window as well as button linkage. It is although utilized to cause the window itself. For illustration, when it occurs to house costs, non environmental factors should be considered. This details will be included into the overall analysis to provide a scale of advantages for a most desired retail. The plan of this structure was implemented with careful notice to detail and effectively utilized the available datasets. Before clink the "price prediction" button, the buyer needs to load the property's location, number of bedrooms (BHK), number of bathrooms, and area(in sqft).

CONCLUSION

The equipped system make use of various factors to make forecast about real estate prices in Bangalore. MultipleMachine Learning procedure were experimented with to find the most efficient model. Among all the task tested, the Decision Tree Algorithm display the primary loss and the topmost R-squared value. The website was constructed using Beaker. Let's display the progress of our design by opening the HTML web runner we situate sooner before launching theapp.py file in the posterior. To evaluate the price of a property, input its square footage, number of bedrooms, bathrooms, and location, and then click on "ESTIMATE PRICE." We have determined the value of an ideal home based on the given knowledge.

REFERENCES

[1]. Number of Retrogression search in insighting House value deviation by Aminah Yusof Md and Ismail Syuhaida .Debating of the IBIMA Vol. 2012(2012), Article ID 383101,9 runners, IBIMA Publishing DOI 10.5171/2012.383101.

[2]. What you actually see might differ from what you actually receive, in accordance with M.A. Babyak A brief, nonspecific beginning to over that fits retrogression- type models. Medical the ground of psychology, 66(3), 411- 421.(3). Atharva Chogle, Priyanka, Akshata Gaud, and Jinal Jain.

[3]. Vaticinating casing prices using information mining systems Vol. 6, Issue 12, December 2017, the International Report of Innovative Studies in facts and Communication Technologies ISO 32972007 pukka .

[4]. Evaluating Zillow foretell Error with Linear Regression & Boosting slants, IEEE 14th universal Workshop on Wireless Ad Hoc & Sensing models, Page(s) 530- 534(4). Sangani Darshan , Kelby Erickson, and Mohammad Hasan.

[5]. Model, A parity On Estimated House cost Using Statistic And Neural Networks, The International diary of Science and mechanics, Research Vol 3, ISSUE 12, December 2014, runner(s) 126- 131, Azme Bin Khamis, Nur Khalidah KhalilahBinti Kamarudin..

[6]. Modelling Home cost vaticinating make use of a Regression Approach and flyspeck Swarming Optimisation Case history Malang, East Java, Indonesia by Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, and Wayan Firdaus Mahmudy. The automation Review of Advanced Computing lores and Applications,Vol. 8,No. 10, 2017,p. 323 –326 .

[7]. Nageswara Rao Moparthi and Dr.N. Geenthanjali," Design and prosecution of hybridization phase determined ensemble model for associating blights using SDLC software criteria ", An transnational conference ordered by IEEE,pp. 268 – 274, 2016 .

[8]. Property sooth saying prices make use of data mining by NiharBhagat, Ankit Mohokar, and Shreyash(8). In October 2016, it was announced in the International diary of Applications of Computing 152(2) 23 – 26.

[9]. Peter B. Luh, Neural Network- Grounded request Clearing Price vaticination and Confidence Interval Estimation With an Improved Extended Kalman Filter Method, IEEETransactions on Power Systems 20(1) 59- 66, March 2005.

[10]. Visit Limsombunchai, Christopher Gan and MinsooLee, House Price vaticination Hedonic Price Model vs Artificial Neural Network, Lincoln University, Canterbury 8150, New Zealand, American Journal of Applied Sciences1(3) 193- 201, 2004..

[11]. Dr. Nageswara Rao Moparthi,, Ch Mukesh,Dr.P.Viday Saga, " Water Quality Monitoring System Using IoT ", An International Conference by IEEE,PP. 109 – 113, 2018.

[12]. Ahmed Khalafallah, Neural Network Grounded Model for Predicting Housing Market Performance, Tsinghua Science & Technology 13(S1) 325- 328, October 2008.