# HOUSE PRICE PREDICTION USING LINEAR REGRESSION ALGORITHM

**Geetika Peddi[1]*, Bhogadi Sai Sri Harsha[2], N.Venkata Koushik[3], Mr.P.Anjaiah[4]**

**Abstract—**
The use of linear regression for predicting housing prices is explored in this study. It makes use of a dataset that includes a variety of important variables, including place, size, and the number of bedrooms. Outliers and missing values are handled using data preparation procedures. To evaluate the model, the dataset is divided into training and testing sets. While regularization strategies manage overfitting, feature selection and engineering techniques isolate relevant predictors. The efficiency of linear regression models in somewhat accurate house price prediction is shown by experimental data. The research results offer real estate stakeholders' knowledge that can be used to estimate home values and guide housing market decision-making.

**Keywords—** Cross validation, Data mining, Price Prediction, Linear regression analysis

[1]*Computer science and engineering, Institute of Aeronautical Engineering,
E-mail:- 19951A0555@iare.ac.in
[2]Computer science and engineering, Institute of Aeronautical Engineering, E-mail:- 17951A05G1@iare.ac.in
[3]Computer science and engineering, Institute of Aeronautical Engineering, E-mail:- 19951A05N5@iare.ac.in
[4]Computer science and Engineering, Institute of Aeronautical Engineering, E-mail:- p.anjaiah@iare.ac.in

***Corresponding Author: -** Geetika Peddi
*Computer science and engineering, Institute of Aeronautical Engineering,
E-mail:- 19951A0555@iare.ac.in

**DOI: -** 10.48047/ecb/2023.12.si5a.0164

## INTRODUCTION

A home is a person's basic need, and the price of a home relies on the characteristics it provides, like parking space, location, etc. Because it is difficult to determine the value of a home based on its size or position in reference to jobs, many residents, whether wealthy or low-income, are worried about house pricing. Purchasing a home is one of a family's most important and essential decisions because it consumes all of their financial resources and, in some cases, blankets them with obligations. Accurately predicting the price of a house can be difficult. Our proposed method enables the precise predicting of property values.

The task of predicting house prices is essential in the real estate sector and has attracted considerable interest from academics and industry experts. Accurately predicting house values offers homeowners, buyers, sellers, and investors useful information that helps with risk management and well-informed decision-making. To address this issue, a variety of machine learning methods have been developed, with linear regression standing out as a core and popular strategy

## I. SYSTEM ANALYSIS

A. *Proposed system*

Regression is the foundation of the suggested approach. To better and more accurately estimate home price trends, this initiative is being suggested. The information used to make the home forecast is gathered from publicly accessible sources. 50% of the dataset is utilized for training during validation, while the remaining 50% is used for testing.

The dataset is divided into a number of subgroups using this approach. At that time, preparation on all of the subsets had been tried; nevertheless, one (k-1) subset remained for evaluation of the prepared model. This tactic emphasizes k times with a different subset each time for the purpose of preparation.

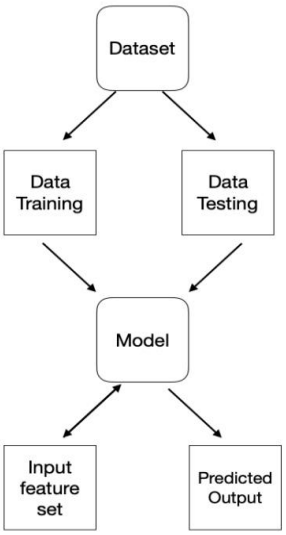



**Fig.**1.Block Diagram

## II. RELATED WORK

### *A. "International Journal of Housing Markets and Analysis""*

This Housing remains to be the most prevalent use of land both locally and globally. In a number of nations, the use of real estate as a financial tool by both individuals and business purchasers has grown significantly in recent years, necessitating further study. The International Journal of Housing Markets and Research will serve as a global platform for the exchange of concepts and knowledge about housing, housing market and their links.

*B. By Chaudhary, A., and Kanth R.A.,"House Price Prediction Using Machine Learning: A Review':*

Housing The latest techniques for prediction home prices, including linear regression, are summarized in this review study. The significance of feature engineering and feature engineering and feature selection in raising prediction accuracy was covered by the authors. They emphasized that linear regression models are transparent and interpretable, making them suited for real estate professional who need clear insights.

*C. Li,Y., & Cui,H., "House Price Prediction Using Multiple Linear and the Decision Tree Models":*

Housing In this study, the effectiveness of decision tree models and multiple linear regression models for predicting housing prices was examined. The writers made use of elements like location, size, and the availability of amenities. The finding showed that despite providing interpretable coefficients to comprehend the impact of each parameter on house prices, linear regression models obtained equivalent accuracy to decision tree models.

## III.METHEDOLOGY

### *A. Data collection*

The systematic process of acquiring knowledge about variables is known as data collection. This aids in the formulation of ideas, the pursuit of answers to different questions, and the evaluation of outcomes. The first step in planning a social event is to collect data, measure it within a structured framework, and use it to assess the outcomes and address important issues. The collecting of data is an essential part of research in all areas of study, including the humanities, commerce, physical and social sciences. While tactics differ depending on the industry, the significance of making sure a decision is accurate and legal does not. For a variety of datasets that

would be pertinent to the goal of our study, it has been tested on Kaggle. This dataset was discovered after a sizable number of datasets were searched. It is an index of house values in California, USA. Less mistakes and differences are present in this assortment for machine learning, which is commonly used.

### *B. Data Cleaning*

To improve the value of data, data cleaning involves finding and correcting mistakes. Data cleansing is carried out using data logic tools. This method is used to locate and change out-of-date records in a database, spreadsheet, or record gathering. It locates the missing information and replaces the jumbled data. The content is altered in order to guarantee precision and clarity. Information cleansing is the process of identifying and removing inaccurate records from a record collection, spreadsheet, or database. It's a method for finding inaccurate information and then replacing it with more correct information. It is altered to ensure that the information is accurate and correct. It is used to ensure a collection's reliability. Finding and eliminating errors is the primary goal of data cleansing, which enables the creation of shifting information estimates. The primary focus should be on choosing the appropriate traits and understanding the relationships between different data ancient oddities, such as trends and records.

### C. *Data Pre-Processing*

The input is changed before entering the software. It is used to convert imperfect data into perfect databases. It's an information retrieval technique that includes rationally organizing unorganized data. The final data that is used for testing and planning is the outcome of data cleansing. The process of transforming useless raw data into something that can be used and created is known as data preparation. Any computer learning technique includes a step called data preparation where the data is altered or saved to make it simpler for the computer to comprehend. The modifications that are made to our data prior to using it for calculation are referred to as pre-handling. Data preparation is a technique for transforming unorganized data into a useful gathering of information. Whenever data is compiled from several sources, it is fundamentally assembled in a way that prevents it from being examined. Real-world data is frequently chaotic, occasionally deficient in certain attributes, and occasionally has an unfavorable structure that makes it illegal to use it in machine learning models. Two of the many tasks that make up data preparation are cleaning and preparing the data for use by a machine learning model. Enhancing the model's precision and competence is another objective.

### D. *Data visualization*

The method of graphically or pictorially displaying data is known as data visualization. It makes it easier to comprehend complex ideas and find brand-new patterns. Many businesses consider data visualization to be cutting-edge pictorial communication. It entails producing and analyzing visual data representations. Information visualization includes measured images, graphs, data patterns, and other tools for clearly showing information. Customers can differentiate between information and evidence and make judgements about it with the help of an effective picture. Access to, understanding of, and usefulness of complicated facts are gradually improved. The job is followed by the reasonable structure standard, such as demonstrating exams or demonstrating causation. Customers may be given duties that are obviously rational, like making judgements or determining the cause. Data depiction requires both skill and focus. While some see it as a part of casual learning, others see it as a tool for developing grounded concepts. In this context, the terms "enormous data" or "Web of things" allude to the vast amounts of data generated by online activity and an increasing number of ground-based devices. For data visualizations, this data presents a wide range of interesting and well-organized issues, including managing, analysing, and disseminating. This task is met by the discipline of data science and its practitioners, called data scientists.

### E. *Exploratory Data Analysis:*

We have opted for the method of multiple linear regression, which calculates the value of the dependent variable based on several independent variables. The projection of the variable in question relies on strength of its correlation with the other independent variables. This measure is commonly known as correlation. Fig 2. Presents a heatmap illustrating the relationships among all the attributes.

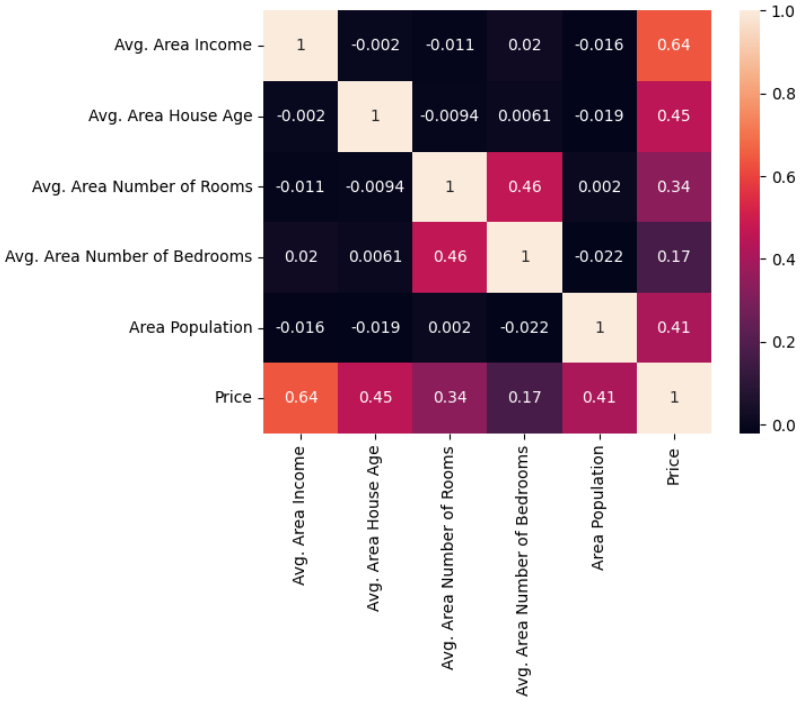



**Fig. 2.** Heatmap on the attributes

While a distplot reveals the distribution of a single variable, a pair plot displays pairwise relationships between variables in a dataset. A pair plot may be used to show how various characteristics relate to one another when using linear regression to predict home prices, and a distplot can be used to show how projected house values are distributed.

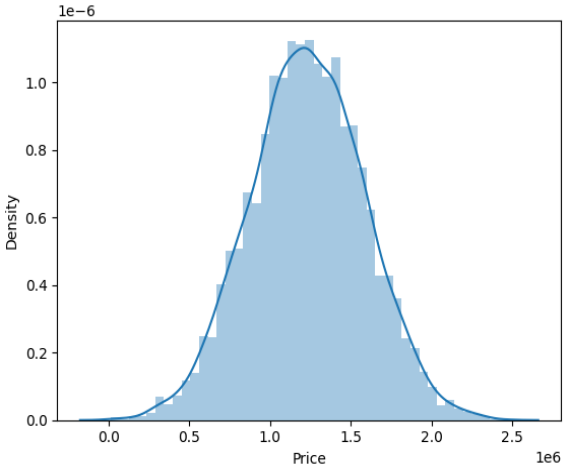



**Fig. 3.** Histogram on Price and Density

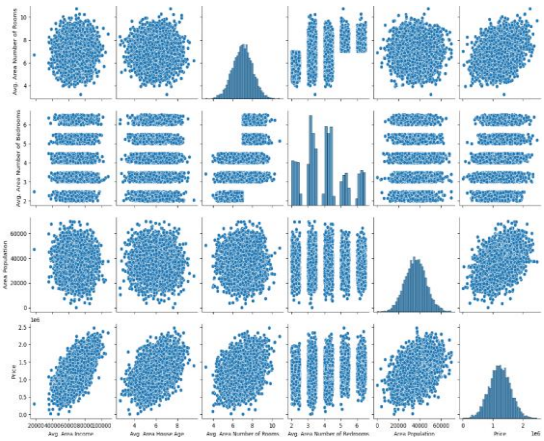


**Fig. 4.** Pair plot of attributes

The x axis is used to represent the independent variables, while the y axis is used to plot the dependent variable. The following is the formula for multivariant linear regression:

$$Y = A + B1X1, B2X2, B3X3..., \text{ and } BnXn$$

Here, y is the dependent variable, and the independent variables are x1, x2..., xn, and the dependent variables are b1, b2..., bn are the coefficient of independent variables.

## IV.RESULTS

Our model's accuracy was assessed using the $R^2$-score as the evaluation metric. The formula is given below:

$$R^2 = \frac{SSR}{SST}$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

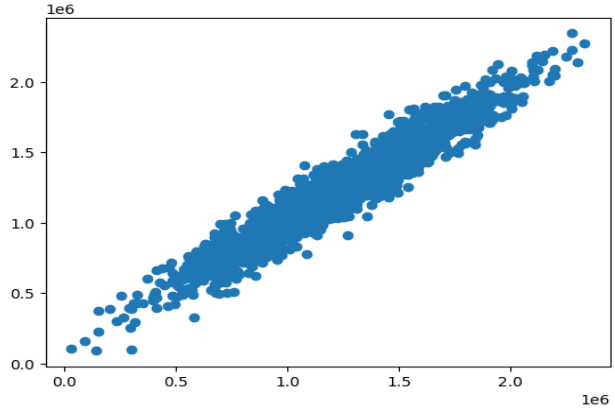
$$SST = \sum_i (y_i - \bar{y})^2$$

Where,

- SSR is Sum of Squared Regression also known as variation explained by the model
- SST is Total variation in the data also known as sum of squared total
- y_i is the y value for observation i
- y_bar is the mean of y value
- y_bar_hat is predicted value of y for observation i

The $R^2$-score, which ranges from 0 to 100%, measures the degree of variance in a dependent attribute that can be predicted from the independent attribute. The two variables are fully correlated when the ratio of the total variance explained by the model to the total variance is 100%, or 1, which indicates that there is no variance at all. As the $R^2$-score's value drops, a regression model's validity gradually deteriorates. The number of data points that fall inside the line drawn by the regression equation is indicated by the $R^2$-score.

$R^2$ is rated 0.91 for our model. Here, it can be shown that our linear regression model can explain 91% of the variance of the dependent variables however the remaining 9% of the changeability is still unaccounted for.




## V. CONCLUSION

Different techniques are used to calculate the commodities' sale values. The sale prices have been determined more precisely and transparently. The general population would gain a lot from this. These findings are obtained using various data mining techniques and the Python computer language. It is important to consider and enhance the numerous factors that influence house costs. We were effective in completing our job using machine learning. First, the process of collecting data is begun. The data is then cleaned to remove as many errors as possible. The analysis of the info is completed after that. The creation of different maps using data visualization follows. In this way, the dissemination of information in various forms has been proven. The model's design and testing are also complete. We discovered that while some categorization techniques worked with our data, others did not. As a consequence, we stopped using the techniques for gathering data on house prices and instead focused on improving the clarity and accuracy of the ones that were still in use. For the purpose of increasing the accuracy of our regression methods. To produce better outcomes, it is critical to increase the programmers' precision and effectiveness. If the results are suspect, residents won't be able to make educated guesses about house acquisition prices. Data visualization was also used to improve accuracy and outcomes. Different techniques are used to calculate the commodities' sale values. The sale prices have been determined more precisely and transparently. The general population would gain a lot from this.

## REFERENCES

1. Siddahant Burse, Dhriti Anjaria, & Hrishikesh Balaji (2021) Housing Price Prediction Using Linear Regression, JETIR ISSN-2349-5162
2. Bhagat, N., Mohokar, A., & Mane, S. (2016). House Price Forecasting using Data Mining. International Journal of Computer Applications, 152(2), 23–26.
3. M. Bhuiyan and M. A. Hasan," Waiting to Be Sold: Prediction of Time- Dependent House Selling Probability," 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, 2016, pp. 468-477, doi: 10.1109/DSAA.2016.58.
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. Vishal Venkat Raman, S. V. (2014). Identifying Customer Interest in Real Estate Using Data Mining Techniques (Vol. 5 (3)). Vellore, Tamil Nadu, India: International Journal of Computer Science and Information Technologies.
6. D. Sangani, K. Erickson and M. A. Hasan," Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting," 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Orlando, FL, 2017, pp. 530-534, doi: 10.1109/MASS.2017.88
7. 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
8. Hu, G., Wang, J., & Feng, W. (2013). Multivariate Regression Modeling for Home Value Estimates with Evaluation Using Maximum Information Coefficient. In R. Lee (Ed.), Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2012 (pp. 69–81). Springer Berlin Heidelberg
9. Mumbai most expensive city in India for expats, ranks 19th in Asia: Survey. (2020). Press Trust of India. https://www.business-standard.com
10. L. Li and K. Chu," Prediction of real estate price variation based on economic parameters," 2017 International Conference on Applied System Innovation (ICASI), Sapporo, Japan, 2017, pp. 87-90, doi: 10.1109/ICASI.2017.7988353
11. Ruben, J. D. (2002). Data Mining: An Empirical Application in Real Estate Valuation. Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, 314–317.
12. 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, Australia, 2018, pp. 35-42, doi: 10.1109/iCMLDE.2018.00017.