Linear regression-based prediction model for real estate properties in India

Harsh Verma
Computer Science and Engineering
ABES Engineering College
Ghaziabad, India
Harsh.20B0101139@abes.ac.in

Amrit Singh
Computer Science and Engineering
ABES Engineering College
Ghaziabad, India
amrit.20B0101106@abes.ac.in

Sandhya Avasthi
Computer Science and Engineering
ABES Engineering College
Ghaziabad, India
sandhya_avasthi@yahoo.com

***Abstract*— The real estate market is one of the most competitive in terms of price, and it is also one of the few that is subject to change on a consistent basis. It is one of the most significant areas where the principles of machine learning are applied to improve and increase the accuracy of cost predictions. The price of a home depends on three factors: its physical condition, its design, and its location. The current method of estimating property prices does not account for fluctuations in market prices or the rate of inflation. The objective of this paper is to provide Indian customers with a forecast of residential prices that takes their needs and budgets into account. To enhance the overall performance of the prediction models, the mean target encoding is also incorporated into the methodology. According to the findings of this study, the estimation error that is produced by linear regression is the lowest and accuracy of the implemented system is 98.6%.**

Keywords*— *Linear Regression, feature engineering, price prediction, accuracy, Decision Tree, Real estate price

I. INTRODUCTION

The rise in real estate prices over the past few years has caught the attention of academics in many different fields. The stability of India's economy, society, and real estate market are all connected [1]. By making predictions and looking at the market, you can find out how stable the real estate market is. On the one hand, it can make it easier for the government to put in place macro-controls on housing costs and keep the Indian economy growing steadily. On the other hand, real estate investors can plan their investments and limit their losses by using house price predictions. China's real estate market started later than those in developed countries like Europe and the United States. China also didn't have enough market experience or policy theory, which led to some bad real estate market policies. Property costs area unit historically calculable supported worth of recently sold properties during a given space and don't take into account factors like close amenities, traffic conditions, proximity to transport and additionally social emotions of the placement. additional recently, the introduction of latest sources of knowledge and ways from the pc vision community, like machine learning, have modified ancient ways of property valuation. Machine learning techniques such as linear regression and grid search CV have become increasingly popular in predicting real estate prices due to their ability to identify complex relationships between variables. The linear regression classification method establishes the correlation between a dependent variable and multiple independent variables, such as location, square footage, and number of bedrooms. Grid search CV is a technique that can be used to identify the best hyperparameters for a model, enhancing its accuracy.

This study aims to provide insights into the potential of using machine learning techniques in the real estate industry, enabling more informed decision-making for buyers and sellers. The results of this study demonstrate the effectiveness of linear regression and grid search CV in predicting real estate prices, highlighting the value of these techniques in the real estate industry.

In this study, we explore the use of linear regression and grid search CV to predict real estate prices. We use a dataset of real estate transactions in a specific area, including features such as the number of bedrooms, square footage, and location. We pre-process the data, including feature scaling, data cleaning, and data splitting for training and testing. Linear regression and grid search CV method is applied to find out best hyperparameters for the model for prediction, and performance is evaluated using mean squared error and R-squared metric. Because of the study's insights into the potential applications of machine learning techniques in the real estate industry, buyers and sellers will be able to make better selections. The results of this study demonstrate the effectiveness of linear regression and grid search CV in predicting real estate prices, highlighting the value of these techniques in the real estate industry.

The first section discusses the price prediction problem of real estate in India. The section II presents a detailed literature review of the existing system for prediction of real estate price. In the third section, proposed framework for prediction model is described. The fourth section will contain implementation-specific experimental data. The paper concluded with a discussion of future research projects and new tourist guide application areas.

II. RELATED WORKS

First, different articles and research papers on the topic are identified and studied. The title of the article is realty value prediction, and it is supported machine learning and neural networks. The publication's description is stripped-down error and also the highest accuracy. Understanding current developments in realty costs and homeownership area unit the topic of the study. Real Estate has become an essential aspect of the 21st century, as it now embodies much more than just a basic need. Not just for individuals trying into shopping for realty however additionally the businesses that sell these Estates. in step with [2] realty Property isn't solely the essential would like of a person however nowadays it additionally represents the wealth and status of someone. Investing in real estate is usually deemed lucrative since property values tend to depreciate at a slower rate. Variations in real estate values have an impact on a wide range of stakeholders, including property investors, financiers, policymakers, and others. Real estate investment appears to be an attractive option for investment, making accurate predictions of real estate values an essential economic indicator.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

According to the paper [1], every organization in the present real estate industry strives to achieve a competitive advantage over its rivals for profitable operations. Therefore, there is a need to streamline the process to deliver optimal results for the average person. [3] projected to use machine learning Associate in Nursing computer science techniques to develop a rule that may predict housing costs supported bound input options. This rule can be applied in business by allowing classified websites to accurately forecast the costs of newly listed properties. By utilizing certain input variables, the website can predict the appropriate and fair market value, thereby eliminating the need for customer input and preventing errors from entering the system. [8]used Jupiter IDE and Google Collab. Jupiter IDE is a web application that works with ASCII text files and allows people to exchange and create documents with LiveCode, visualizations, equations, and narrative text. It includes tools for knowledge enhancement, knowledge transformation, numerical value simulation, modeling statistical exploitation, knowledge image, and machine learning tools.[7] designed a system that may facilitate individuals to grasp on the point of the precise value of realty. User will offer their needs in step with that they'll get costs of the required homes User can even get the sample arrange of the house to induce a reference for homes.[4] used an information set of one hundred homes with many parameters. we have used fifty % of the information set to coach the machine and fifty % to check the machine. The accuracy of the results has been confirmed, and the testing has been conducted using various parameters. As described in [9], the study employed fundamental machine learning algorithms, including decision tree classifier, decision tree regression, and multiple regression to the mean, and executed them with the Scikit-Learn machine learning tool. The purpose of this project is to assist users in predicting both the supply and prices of homes in a given town. Machine learning algorithms were utilized to predict house prices, as detailed in [5]. A stepwise process was followed to analyze the dataset, and after inputting feature sets into four different algorithms, a CSV file containing the predicted house prices was generated. However, as noted in [6], it is important to use a combination of models, as a linear model may suffer from high bias (underfitting), while a complex model may suffer from high variance.

TABLE I. SUMMARY OF PRICES PREDICTION MODELS

Study	Paper Description			
	Objective	***Data Set***	***Method***	***Result***
[1],(2020)	To forecast house prices using machine learning	Dataset of house sales in Pune, India from 2010 to 2019	Linear regression, Random Forest, XGBoost, SVM	R-squared values is 0.91
[2],(2018)	To develop a model for house price prediction using machine learning	Dataset of house sales in Pune, India from 2010 to 2017	Linear regression, Random Forest, K-Nearest Neighbors	R-squared value of 0.86
[3],(2019)	To develop a PropTech solution for proactive pricing of houses in classified advertisements in the Indian real estate market	Data on house prices and features in Indian real estate classified ads	Machine learning algorithm using decision tree regression	R-squared value of 0.83
[4],(2019)	To develop a model for house price prediction using machine learning	Dataset of house sales in Mumbai, India from 2015 to 2018	Multiple linear regression, Support vector regression, Artificial neural network (ANN)	R-squared value of 0.87
[5],(2021)	To develop a model for property price prediction using machine learning	Dataset of property transactions in Hong Kong from 2014 to 2019	Multiple linear regression, Random Forest, Gradient Boosting, Neural network	R-squared value of 0.82

III. PROPOSED FRAMEWORK AND METHODOLOGY

We are using different algorithm in this project as the proposed algorithm and various methods to clean and process the data. Among these algorithms we have checked the best model to predict the data. These algorithms are discussed below:

A. Linear Regression

A continuous output variable can be predicted using the statistical method of linear regression using one or more input factors in machine learning. It operates on the assumption that there is a linear connection between the input and output variables, and it seeks out the best-fit line that can forecast the output value for any given input value.

The equation for a simple linear regression model with one input variable (x) and one output variable (y) can be expressed as in (1):

$$y = b0 + b1 * x \quad (1)$$

In linear regression, the predicted output variable (y) is determined by the y-intercept (b0) and the slope of the line (b1), which indicates the change in y for a one-unit increase in the input variable (x). The primary goal is to obtain the b0 and b1 values that minimize the discrepancy between the predicted output and the actual output values in the training data. This is typically achieved using the least squares regression technique, which involves identifying the values of b0 and b1 that minimize the sum of the squared differences between the predicted and actual output values.

B. Decision Tree Regression

Decision tree regression is a machine learning algorithm that utilizes one or more input variables and a continuous output variable to make predictions. This supervised learning method partitions the data into subsets based on the input variables in a recursive manner, and fits a basic model like the mean or median value to each subset.

At each node, the decision tree algorithm selects the optimal variable to divide the data based on a measure of the

homogeneity or purity of the resulting subsets, resulting in the creation of the decision tree. The goal is to create a tree that maximizes the homogeneity of the subsets and minimizes the variance of the output variable within each subset.
The equation for a decision tree regression model is not as simple as in linear regression, since it involves multiple decision rules and model parameters. However, the basic idea is to use a series of if-then statements to predict the output variable based on the input variables.

C. Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression is a type of linear regression algorithm used in machine learning to reduce the complexity and improve the accuracy of a linear regression model. It works by adding a penalty term to the objective function that the algorithm minimizes, in order to discourage large coefficients for the input variables.

The equation for Lasso regression with one input variable (x) and one output variable (y) can be expressed as in (2):

$$y = b0 + b1 * x + e \quad (2)$$

where y is the predicted output variable, b0 is the y-intercept, b1 is the coefficient for the input variable x, and e is the error term.

The goal of Lasso regression is to find the values of b0 and b1 that minimize the objective function, subject to the constraint that the sum of the absolute values of the coefficients is less than or equal to a certain threshold. This is typically done using optimization techniques such as coordinate descent or gradient descent. Lasso regression can be used to improve the accuracy of linear regression models, especially when there are many input variables and some of them are irrelevant or redundant. It can also be used for feature selection and model interpretation, since it tends to set some of the coefficients to zero and thereby identify the most important input variables.

D. K Fold Cross Validation

K-fold cross validation is a popular approach in machine learning for assessing model performance and tuning hyperparameters. The method involves dividing the available data into K subsets, or folds, of equal size. One of the folds is used for testing the model, while the remaining K-1 folds are used for training. This process is repeated K times, with each fold used once for testing, and the overall performance is determined by averaging the results of each fold.
K-fold cross validation helps to improve the reliability of model evaluation by reducing the variability of the evaluation metric. It also helps to prevent overfitting of the model by providing a more realistic estimate of the model's performance on new, unseen data.
The value of K can be chosen based on the size of the available data and the computational resources available. Common values of K are 5 or 10, but other values can be used depending on the problem.

E. Grid Search Algorithm

The grid search algorithm is a method in machine learning that is utilized for hyperparameter tuning, which aims to determine the optimal combination of hyperparameters for a given model. Hyperparameters are settings that are specified before training begins, and are not learned by the model during training. The grid search technique operates by defining a grid of possible hyperparameter values to explore, and then evaluating the model's performance on a validation set for each hyperparameter combination in the grid. The validation set is typically a subset of the training data that is held out for the purpose of evaluating the model's performance. The grid search algorithm exhaustively searches over all combinations of hyperparameters in the grid, and returns the combination that results in the best performance on the validation set. The performance metric used for evaluation can vary depending on the problem, but typically includes measures such as accuracy, precision, recall, F1 score, or mean squared error.

IV. IMPLEMENTATION

A. Problem Statement

Create a model to estimate the worth of homes.

B. Data

The most crucial aspect of a machine learning project is the data, which requires special attention. The results of the project can be significantly influenced by various factors related to the data, such as the source, format, consistency, and presence of outliers. Therefore, several questions must be addressed during this stage to ensure the efficacy and appropriateness of the learning algorithm. In order to obtain, clean, and transform the data, several sub-steps are necessary. To gain a deeper understanding of how these steps have been implemented in the project and their significance for the machine learning phase, we will review them in detail.

C. Frontend

The side is made from easy hypertext mark-up language. To receive Associate in Nursing calculable rating, the user could fill-up the shape with the amount of sq. feet, BHK, bathrooms, and placement and click on the 'ESTIMATE PRICE' button. we have used Flask Server and organized it in python. It takes the shape knowledge entered by the user and executes the operate, that employs the prediction model to calculate the projected value in lakhs of rupees

D. Project Architecture

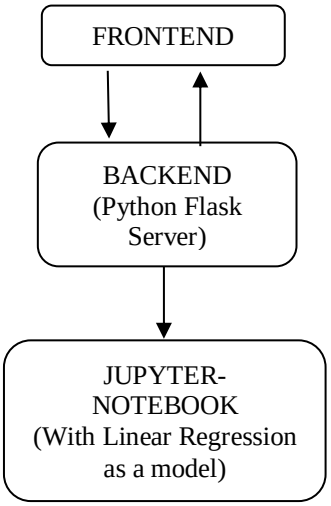

Fig. 1. Architecture of Application

E. Data Science

The first stage begins with knowledge science, that take the information set and do some vital cleansing the data to ensure that it provides reliable prediction throughout. The Jupyter notebook "Real-Estate-Price-Prediction.ipynb", is wherever all the science work is hold on. In terms of information cleansing, our dataset desires a big quantity of effort. In fact, most of the notebook is devoted to knowledge cleansing, during which we tend to eliminate empty rows and take away superfluous columns that may not aid in prediction.

The final stage is to modify outliers. Outliers' area unit abnormalities that do huge harm to knowledge and prediction. Finally, the initial dataset of over 13000 rows and nine columns is reduced to concerning 7000 rows and five columns.

F. Machine Learning

The ensuing knowledge is fed into a machine learning model. to search out the best procedure and parameters for the model, we are going to largely use K-fold Cross-Validation and also the GridSearchCV approach. It seems that the simple regression model produces the most effective results for our knowledge, currently we've to transfer our mythical being and pickle file to our flask server to convert it into python objects and hook up with the frontend. Experimental setup

V. Experimental Setup

A. Step to Create Model

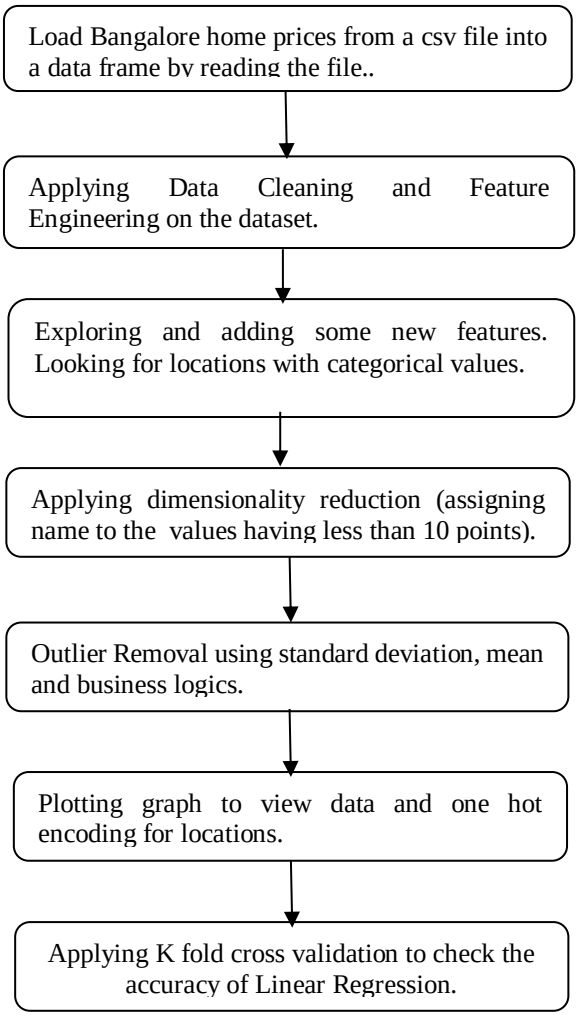

Fig. 2. Flow Diagram for creating Model

B. Tools and Technologies Used

1) Flask

When a Flask application is run, it starts a server that listens for incoming HTTP requests. When a request is received, Flask processes it and returns an HTTP response. To run a Flask application as a server, the developer typically creates a Python file that defines the application and includes the necessary Flask modules and dependencies. The application can then be started by running the Python file using a command such as python app.py in the terminal.

2) Python

Python is a versatile and widely adopted high-level programming language, which is interpreted and suitable for a variety of purposes such as web development, data analysis, scientific computing, artificial intelligence, and machine learning. One of the key features of Python is its clean syntax, dynamic semantics, and robust libraries, which make it a preferred choice for both novices and experts. Python is renowned for its readability, user-friendliness, and extensive community support.

3) HTML

HTML, which stands for Hypertext Markup Language, is the principal markup language used to create web pages and web applications. Its main purpose is to structure and organize content on the internet and serves as the fundamental building block for all websites. HTML uses tags to define elements such as headings, paragraphs, images, links, and tables. HTML also supports multimedia, interactive forms, and scripting languages such as JavaScript.

4) CSS

Cascading Style Sheets (CSS) is a language used to describe how HTML or XML documents should be presented to users, by defining the style, layout, and formatting of elements on a web page. Layout, color scheme, font selection, and other visual elements of a web website are all defined by CSS. A website can adjust to various screen sizes and devices by using responsive web design, which can be created using CSS. The user experience on a web page can be improved by using CSS, which also enables animations and transitions.

5) Anaconda

Anaconda is a well-known open-source distribution of programming languages, including Python and R, that are utilized for scientific computing, data analysis, and machine learning. It comes with several pre-installed libraries and tools, including Jupyter Notebook, pandas, NumPy, Matplotlib,

6) Jupyter Notebook

Jupyter Notebook provides a convenient platform for analyzing data and creating visualizations using Python, R, and other programming languages. Jupyter Notebook can be

used for building machine learning models, exploring data, and evaluating model performance. Jupyter Notebook is an ideal environment for cleaning and preprocessing data, which is a crucial step in any data analysis or machine learning project**.**

VI. Result

The result of the project is produced in the form of the website and whose various snapshots are attached herewith in the result.

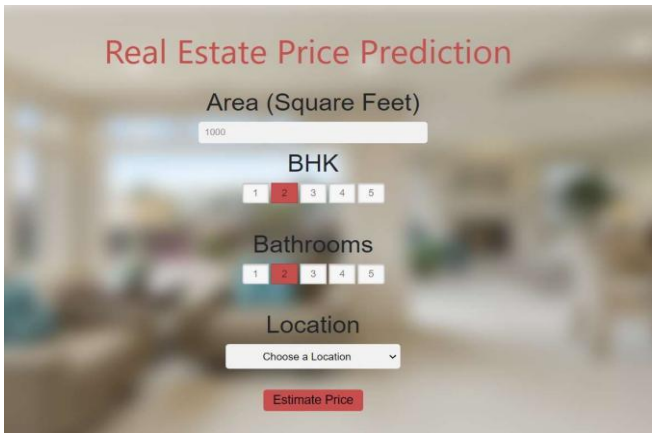

Fig. 3. Interface for taking inputs

As it is clearly visible in Fig. 3. that the output is asking the user for various fields for choosing the suitable house with the best and accurate price.

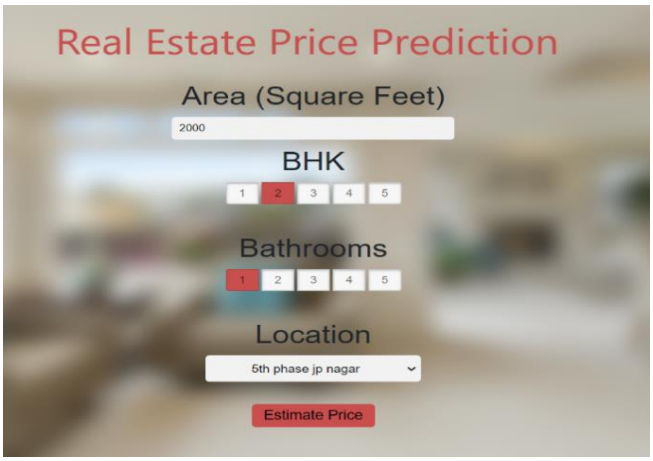

Fig. 4. Inputs filled by user

After filling the appropriate information in Fig. 4 and by clicking the "ESTIMATE PRICE", a new screen for showing the price will appear which is shown as in Fig. 5

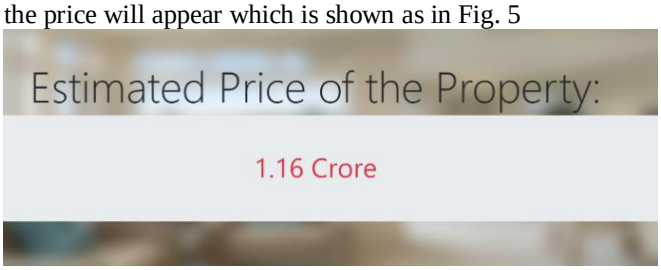

Fig. 5. Predicted Price

Hence the price is shown.

VII. Conclusion

The goal of our proposed model is to help people purchase homes and real estate at fair prices and avoid being misled by unscrupulous agents motivated solely by profit. By analyzing various characteristics of a given dataset, our system predicts the appropriate property price. We tested different Machine Learning algorithms and found that Linear Regression Algorithm performed the best in comparison to others. With the help of various software technologies and systems, we can enhance the accuracy of these predictions by incorporating attributes such as the environment, market trends, and other relevant variables related to real estate. These predictions can be saved in a database, and an application can be developed for the general public. This would enable individuals to make informed investment decisions and minimize potential risks associated with real estate investments.

References

[1] Kuvalekar, Alisha and Manchewar, Shivani and Mahadik, Sidhika and Jawale, Shila, *House Price Forecasting Using Machine Learning* (April 8, 2020). Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST) 2020

[2] Neelam Shinde, Kiran Gawande, "*Valuation Of House Prices Using Predictive Techniques*",International Journal of Advances in Electronics and Computer Science, Volume-5, Issue-6, 2018.

[3] Putatunda, S. (2019). PropTech for *Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market.* arXiv preprint arXiv:1904.05328.

[4] Atharva Chouthai, Mohammed Athar Rangila , Sanved Amate, Prayag Adhikari, Vijay Kukre , "*House Price Prediction Using Machine Learning*" , International Research Journal of Engineering and Technology(IRJET), Vol:06 Issue: 03, 2019.

[5] Ho, W. K., Tang, B. S., & Wong, S. W. (2021). *Predicting property prices with machine learning algorithms*. Journal of Property Research, 38(1), 48-70.

[6] Akshay Babu, Dr. Anjana S Chandran, "*Literature Review on Real Estate Value Prediction Using Machine Learning*", International Journal of Computer Science and Mobile Applications, Vol: 7 Issue: 3, 2019.

[7] Mr. Rushikesh Naikare, Mr. Girish Gahandule, Mr. Akash Dumbre, Mr. Kaushal Agrawal, Prof. Chaitanya Manka, "*House Planning and Price Prediction System using Machine Learning*", International Engineering Research Journal, Vol:3 Issue: 3, 2019.

[8] Bindu Sivasankar, Arun P. Ashok, Gouri Madhu, Fousiya S, "*House Price Prediction*", International Journal of Computer Science and Engineering (IJCSE), Vol: 8 Issue: 7, 2020.

[9] M Thamarai, S P Malarvizhi, " *House Price Prediction Modelling Using Machine Learning*", International Journal of Information Engineering and Electronic Business (DJIEEB), VoL12, No.2, pp. 15-20, 2020. DOI: 10.5815/ijieeb.2020.02.03.