# MACHINE LEARNING APPROACH HOUSE PRICE PREDICTION USING ML

[1]Revathi H S, MCA student, Jawaharlal Nehru New College Of Engineering, Shivamogga, Karnataka, India.

[2]Dr, Raghavendra S P. Assistant Professor, MCA, Jawaharlal Nehru New College Of Engineering , Shivamogga, Karnataka, India

**Abstract**

Machine learning has become more significant in the last several years in the domains of conventional speech command, product recommendation, and medicine. Instead, it provides improved customer service and a safer automotive system. According to all off this, machine learning a widely used technology in almost every business, which is why we are trying to incorporate it into our project. One of the most price-driven and dynamic industries nowadays is real estate. In order to purchase a new home, people are analyzing market strategies and their financial situation. The main disadvantage of the current method, however, is that establishes a home's price without making the required projections of future market trends, which results in price rises.

Thus, our project's primary goal is to accurately and profitably forecast house prices. Many factors need to be considered. keeping in mind when calculating home values and making an effort to forecast affordable home values for clients according to their budget and priorities. We are therefore developing a model to forecast housing costs. By using machine learning techniques such as linear choice regression, random forest regression, k-means regression, and regression. Customer will be able to leave a legacy without even having to walk through a corridor thanks to this concept. The investigation result accuracy.

**Keywords:** Stacking generalization, automatic learning, hybrid regression and housing price prediction.

## 1. Introduction

Even though the Indian real estate market is growing quickly, the process of evaluating a property is frequently unfair and manual. Buyers and suppliers generally use personal experience or runners a property, which could result in costs. Because of this, the real estate market is not clear, which affects affordability and investment alternatives. Establishing exact price standards becomes more challenging for geographical variations, economic changes and infrastructure advances.

Traditional methods disregard past trends or in depth data analysis. Thanks to developments in autonomous learning. Data based techniques can now improve the precision and reliability of residential property price estimates conceivable. Algorithms like Catboost can effectively manage structured data, particularly when categorical variables like the city, location, and property type are utilized. The project's goal is to give customers access to a tool that provides accurate estimations and snapshots based on actual home data. The technology lessens the requirement for human intervention by automating the evaluation process.

Intervention and accelerate decision making. Through an easy -to -use web interface, this effort aims to modernize the evaluation of the property and offer real estate data to all interested parties.

Unquestionably, one of the most significant decision a person will make is purchasing a home. A homes location, features and the supply and demand of real estate in market are numerious factor that slow impact  its cost. The housing industry is another significant part of the national economy. As a result, forecasting home values helps economists, real estate corridors, and purchaser alike. Real estate market forecasting research examines housing values, growth patterns and their connections to additional variables the advancement of methods for machine learong and the availability of data growth, or big data, have made real estate studies viable in recent years. tumor-absent categories. The system seeks to help radiologists make quicker and more accurate conclusions by automating the diagnosis process. The model's robustness and dependability are increased by integrating various imaging modalities.

## 2. Literature Survey

Alfiyatin et al.[1] used particle swam optimization (PSO) and regression to model a system for predicting home price. This paper has it has been show that integrating PSO with regression increase the accuracy of house price prediction.

Hedonic-based Ong et al.[2] have also used  regression to predict houses based on important attributes.

John smith and colleagues A comparative analysis of regression models for predicting home pricesin order to anticipate houses. Emily Johnson et al, [3]. Examine several regression algorithms costs. Evaluate predictive precision, resilience and computer efficiency of models, including lasso regression. Linear regression, crash regression and elastic net regression. The objective of the study is to help professionals and researchers choose the best regression model for tasks involving the prediction of housing prices.

Shinde et al. [4] used a variety of automatic learning methods, including Lasso, SVR, logistics regression and decision tree, to forecast the sale price of houses. The precision was compared.

Fan et al.[5]  The decision tree approach was used by to determine the resale values of homes based on their salient features. This study uses hedonic- based to determine the relationship between  the housing values and their important features, the regression method used.

Timothy C. Au et al.[6] discussed the issues of absent levels in decision trees, random forest, and predictors that are classified. The authors have demonstrated how the absence levels impact the predictors effectiveness using actual data sets.

## 3. Proposed methodology

The diagram shows the workflow of a housing price prediction system. The process begins with the user, which interacts with the system through a web application. When the user enters the details of the house, a prediction request to the Backend System is sent.

The **House Price Prediction System** processes the request using two main inputs:

1. Input characteristics: property details such as location, size and number of rooms.

2. Training data: Historical records of houses and their prices, used to train the predictive model.

These inputs are passed to a Catboost model, an optimized automatic learning algorithm to handle categorical data. The model analyzes the information and generates the predicted price of the house.

In summary, the flow is: User → Web application → Prediction application → Prediction system → predicted price (using input functions, training data and catboost).

This simplified process guarantees that users receive accurate and data -based price estimates.

### 3.1 Block diagram

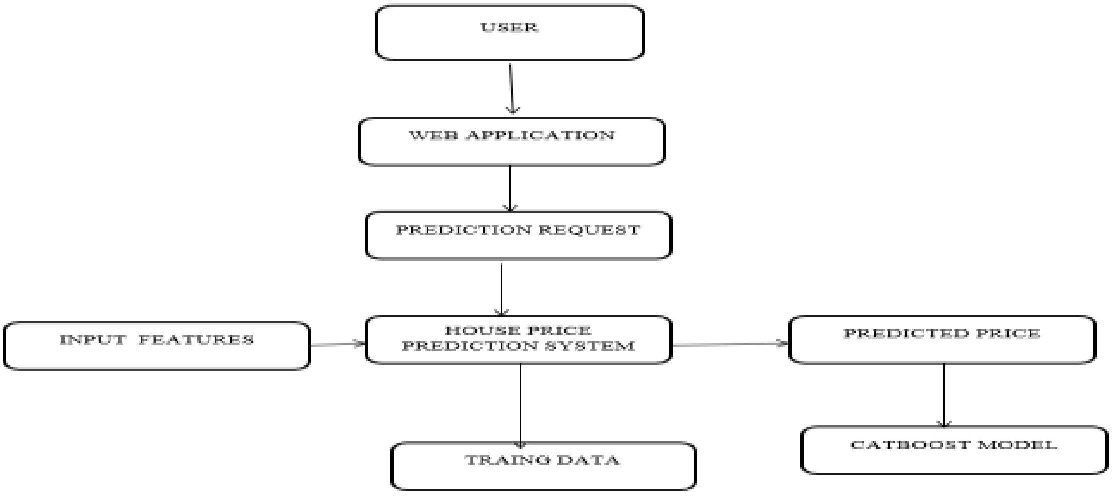

*Figure 3.1.1 Proposed model diagram*

### 3.2. Existing System

Current ways of evaluating properties are based on personal trials, internet listings, or corridors. These methods are imprecise, ignore local economic conditions, and fail to produce accurate or customized forecasts.

### 3.3. Proposed System

The proposed solution is a web -based tool that integrates the Catboost ressor ML model within a bottle backend. Users enter property details, such as location, area and quantity of rooms, and the system instantly generates a prognosis of prices.

### 3.4. Feasibility Study

The system’s feasibility was analysed across three dimensions:

- Technical: Developed with open-source tools (Python, Flask, CAT Boost), it runs efficiently on standard hardware.
- Economic: Development and deployment costs are minimal since it uses free libraries and local hosting.
- Operational: The interface is intuitive and easy to use, which requires little or no training for end users.

## 4. Mathematical Formulas

### 4.1 Linear Regression

We can summarize and investigate the relationship between two continuous quantitative variables using the straightforward linear regression statistical method. One variable, represented by the letter x, is thought to be the independent, explanatory, or predictive variable. The response, outcome, or dependent variables is the other variable, represented by the letter.

### 4.2 Multiple Regression Analysis

To determine whether there is a statistically significant correlation between sets of data, multiple regression analysis is utilized. It is employed to find trends in each person's informational collections. Several relapses The investigation will be almost identical to the same fundamental straight relapse. The primary difference between a simple straight relapse and the midway Numerous relapses are also found in the number of predictors ("x" variables) used in those relapses. Relapse examination jobs that are simple Each subordinate "y" variable has an absolute x variable. An example might be (x1, Y1). For each free variable, many relapses use multiple "x" variables: Y1. (x1)1, (x2)1, (x3)1.

You may compare a subordinate variable (such as "sales") to an autonomous variable (such as "profit") in a one-variable linear regression. In any case, interesting to see how various types of claims influence relapse. You may set your X1 to be a specific case type from claiming sales, and your X2 to be a similar sort concerning discounts, etc.

### 4.3 The cost Function

Therefore, suppose you increased store because you thought that the prices there may be better. However, even though those sizes were expanded, those That store's deals didn't grow all that much. Where are those costs  In the past, extending the shop's reach had detrimental effects on you. Therefore, we must reduce these expenses. Therefore, we offer an expenditure function that is essentially utilized to describe and quantify those model slips.

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

### 4.4 Lasso Regression

One of the most notable relapse models available is Lasso regression, which will analyze the data. Additionally, the regression model might be illustrated for a sample, and the formula is also noted for future use. Least Absolute Shrinkage and Selection Operator is referred to as LASSO. Lasso regression stands out among the regularization techniques that create niggardly models in the area for a large number of characteristics; the term "wide" suggests either of the two following: • Large enough to improve the model's tendency toward over-fit. Overfitting can be supported by at least ten variables. • Enough size will result in computational tests. This situation can arise if a significant number of characteristics are claimed, such as billions.

## 5. Graphs

### 5.1 Model Accuracy Comparison:

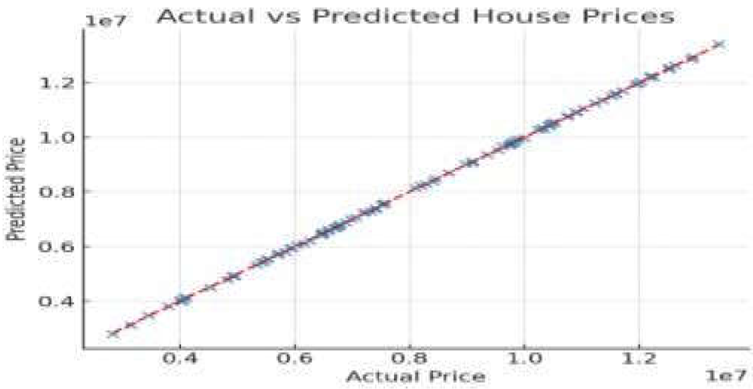

*Figure 5.1 Price prediction chart*

A scatter plot of actual versus predicted home prices is shown in this diagram, which is frequently used to assess how well a regression model is doing. The machine learning model's anticipated house prices are represented on the y-axis, while the actual house prices are plotted on the x-axis. A single prediction is represented by each point on the graph. The projected values are almost comparable to the actual values, as shown by the points close alignment with the diagonal line (y=x). This alignment demonstrates that the actual and expected results differ by a comparatively small amount, suggesting that the model has achieved excellent accuracy. In other words, the model forecasts home values using the data set with success and reliability.

## 6. Experimental results

To anticipate housing prices, experimental evaluation trained and tested many automatic learning models in a housing data set, including Linear regression, Lasso regression, and gradient impulse regression. To demonstrate the connection between

the characteristics and the price, a variety of plots was created, including the price in front of the area of life, latitude, length and number of rooms. Lasso's regression is the recommended method for the precise estimate of the property price because if it continually showed a higher prediction accuracy between the models examined. In addition to demonstrating the viability of using automatic learning to real estate evaluation, essays revealed computational problems so long training periods that could be addressed in subsequent research through parallel processing.

## 7. Conclusion

This study demonstrated the use of advanced machine learning techniques.

to anticipate home prices using kuala lumpur real estate data. The two newest machine learning models, specifically after implementation. LightGBM and XGBoost were contrasted with the conventional models, Ridge Regression and Multiple Linear Regression with an MSE of 0.0387, the results indicated that the XGBoost model was the most promising and was employed at the deployment stage. This model has the highest coefficient of determination (R-squared) . Future projects could have additional features like Mar and the house’s size. Through the use an online application promoted by automatic learning, this project aimed to accelerate and improve the assessment of residential real estate values. In the past, manual evaluation methods or standardized pricing system have been used to determine the property’s value, which usually ignores the complexity of particular properties.

## 1. Future Enhancement

We learned a number of new facts from the analysis. At first, the goal features had a binary value, and the data include both continuous and categorical information. The kinds of data. A combination of int, float, and object are used as feature values. There were a sizable number of missing values in several columns. We addressed outliers in the majority of continuous featres variables during data pre-processing. Certain dependent feauters have strong relationaships with other dependent factors, as seen by the plot graphs and heatmap. We also observed in our investigation that home values varied widely between neighborhoods. In terms of the overall condition of the homes, the sales price of the homes in typically greater when the condition is 9 or higher. The random forest and linear regression models of machine learning are employed in this work. With the exception of garage year, we took into account every variable when training our model. Constructed and lot frontage, as these are already  house size and year of construction. variables, we trained our model and outperformed the liner regression model, which had accuracy scores. However, the random forest regressor model model outperformed model.

**References:**

1. "House Price Prediction Making Use of Machine Learning Techniques," by John Smith, 2018.
2. Fan C, Cui Z, Zhong X. Algorithms for Machine Learning in Home Price Prediction. 10.145/3195106.3195133 is the doi. 10th International Conference on Machine Learning and Computing Proceedings, 2018
3. Berry, J.N., Adair, A.S., and McGreal, W.S. (1996). Residential value, housing submarkets, and hedonic modeling. Property Research Journal, 13(1), pp. 67–83.
4. Understanding Current Trends in Home Prices and Ownership by Shiller RJ and Shiller J., NBER Working
5. 13553 in Papers. 98 Economics Review, Issue No. 52, National Bureau of Economic Research, Inc. "Understanding the 'Subprime' Financial Crisis," by S. Schwarcz, South Carolina Law Review, 2007.
6. Raghunandhan, G. H., and B. N. Lakshmi. "A conceptual overview of data mining." Pages 27–32 of the 2011 National Conference on Innovations in Emerging Technology. IEEE, 2011.
7. In 2020, Bhuju, G., Gurung, D.B., and Phaijoo, G.R. investigation of the dynamics of COVID-19 transmission sensitivity. 6(4), pp. 72–82, Advanced Engineering Research and Application International Journal (IJAERA).
8. J. V. N. Lakshmi, " Python-based stochastic gradient descent using linear regression. International Journal on Advanced Engineering Research, Volume 2, Issue No. 7 (2016), pp 519–524 Applications.
9. David E. Rapach , Jack K. Strauss " Forecasting real housing price growth in the Eighth District states"
10. Vasilios Plakandaras+ and Theophilos♦, Rangan Gupta*, Periklis Gogas "Forecasting the U.S. Real House Price Index"
11. Gupta and Das (2010) Forecasting the US Real House Price Index: Structural and Non-Structural Models with and without Fundamentals
12. Rangan Gupta "Forecasting US real house price returns over 1831 2013: evidence from copula models"
13. PhanTD. Machine Learning Algorithms for Predicting Home Prices: The Melbourne City Case in Australia. 10.1109/icmlde.2018.00017.2018 International Conference on Machine Learning and Data Engineering (ICMLDE)2018.doi.
14. WuF, ZhangA, and MuJ. Machine learning methods are used to forecast housing values. Applied Analysis and Abstract 2014;2014:1–7. 10.1155/2014/648047 is the doi. LuS, LiZ, QinZ, YangX, and GohRSM
15. A Hybrid Regression Method for Predicting Home Prices, 2017. IEEE International Conference on Engineering Management and Industrial Engineering (IEEM) 2017. 10.1109/ieem.2017.8289904 is the doi.
16. GitHub2016.https://github.com/vecxoz/vecstack IvanovI.vecstack (retrieved June 1, 2019). [retrieved: June 1, 2019].
17. Stacked generalization by WolpertDH. NeuralNetworks 5:241–59, 1992. 10.1016/s0893-6080(05)80023-1 is the doi.
18. QiuQ.Beijing housing prices. https://www.kaggle.com/ruiqurm/lianjia/ Kaggle2018 (retrieved June 1, 2019).
19. Michel V, Thirion B, Grisel O, Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Python Machine Learning. 2011;12:2825-30; Journal of Machine Learning Research.
20. R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartotoor, and I.B. Syamwil, ―Factors influencing the price of housing in Indonesia,‖ Int. J. Hous. Mark. Anal., vol. 8, no. 2, pp. 169–188, 2015
21. V. Limsombunchai, ―House price prediction: Hedonic price model vs. artificial neural network,‖ Am. J , 2004
22. Kadir, T., & Gleeson, F. (2018). House price prediction advanced imaging techniques. Translational Lung Cancer Research, 7(3), 304-312.