A Comparative Study of Machine Learning Approaches to House Price Prediction

Sakruthi Avirineni

Computer Software Engineering Department, San Jose State University,
San Jose, California, United States of America
sakruthi.avirineni@sjsu.edu

Abstract - Predicting the house price in the current real estate industry is very challenging and it is more difficult for the lower and middle class people who don't have enough budget. Our research helps in predicting the property value by considering the different economic factors. For identifying the best way of predicting house value and for getting better accuracy we have used various regression and machine learning models like SVR (Support Vector Regression), Random Forest Regression, XG Boost regression and Decision Tree Regression. The models are developed utilizing a variety of factors that affect housing costs, including location, amenities, amenities, and interest rates.Due to the rising cost of living in big cities and in such competitive housing markets, estimating the final cost of a home might be dangerous. This research article focuses on resolving this problem by outlining a technique for precisely forecasting home prices. The price of a property is determined by the analysis of customer wants and financial resources. The research paper helps to provide accurate price predictions for the clients and serves the builder in setting the selling price that aligns with the customer's needs.The findings have important suggestions for those looking for cheap home alternatives and builders hoping to meet consumer demand. Informed decision-making, financial preparation for potential purchasers, and competitive pricing techniques are all made possible by accurate house price estimates. The research makes a contribution to the area of real estate, by offering feasible solutions to the difficulties experienced by lower- and middle-class people in the housing market.

Keywords – house price, support vector regression, XG Boost, decision tree regression, random forest regression.

I. INTRODUCTION

Shelter is among the most crucial need for human life. A person can feel safe and comfortable in their home [4]. Everyone needs their own house, yet the present real estate market experiences rise and decline in property values at various times. The majority of businesspeople seek to spend their money in any industry in the current world in order to increase their profits. Money can be invested in different ways, includes the stock market and buying property for the

family. Many people wish to utilize their money to purchase a home. As the population grows, so the demand for real estate also increases. As a result, making it easier for people to buy their first home on a budget is getting more difficult. Because real estate is a large industry, millions of dollars and sold hundreds of properties and acquired each year. When purchasing a property in a city like Bengaluru, a buyer takes into account a number of aspects, including location, climate, size of the lot, distance from parks, hospitals, and power plants, as well as the kitchen, balcony, bedroom, bathroom, and, most significantly, the price of the home.

It is very difficult for engineers to analyze and predict the house price. The paper focus on various machine learning algorithms and Regression Techniques [3] using several factors for predicting the house price. In this study, tried to build a house prediction model using machine learning models for the dataset that is accessible on the Machine Hackathon platform. This paper discusses the various prediction models, the ordinary least squares model, Lasso and Ridge regression models, the Support Vector Regression model, and the XG Boost Regression model.

The structure of the paper is divided into the various sections. Section 2 literature review addresses the previous research on the subject and also the solutions presented by different authors. Section 3 project justification, which explains the importance of the research paper. Section 4 technical aspects of the paper are discussed in this section which includes the description of the dataset and exploratory data analysis of the dataset, the comparative study of regression models, implementation, limitations, system architecture, and and description of the models for solving the problem. Section 5 summarizes all the models, finds which model is best based on the accuracy of the different models, and recommendations for future work. Section 6 includes the 10 peer-reviewed papers that are considered as a reference for this paper.

II. LITERATURE REVIEW

Researching a prospective home before making a purchase is quite typical in every nation. As time goes on, the price of a property fluctuates, so figuring out what influences it might assist forecast its future value. There are numerous various elements that influence the price of a home, according to a study. Different criteria were employed by each study report to forecast the price of a home. Five variables have been identified as determining home prices according to a article produced by N. S. A. Lati and her colleagues [8]. Foreign Direct Investment (FDI), Gross Domestic Product (GDP), interest rate, unemployment, and inflation are five of these elements. Another study by Ong Tze San established the significant influences of GDP, population, and RPGT on changes in housing prices. GDP, loan rate, and local population seem to be the common components for the majority of research articles. After analyzing data from Bursa Malaysia, Lim Sze Yoong and his team's article further showed that the three factors described above had a significant impact on home prices.

However, with machine learning, in addition to the characteristics mentioned in the preceding paragraph, other variables are used to estimate house prices. A. P. Singh and his colleagues [1] included a variety of factors in their prediction model in a study report, including population, unemployment rate, labor costs, and more. These factors include the size, the distance to the nearby city, and the distance to the nearest school.

There are several methods for forecasting house prices in machine learning. Regression modeling is used by one of them. Regression analysis was used to forecast the property price in a case study by J. Manasa and her colleagues [3]. Prediction of house price is done by linear regression, and they achieved good results as long as they had access to enough data. T. M, S. Masrom, and Z. Li applied the random forest , decision tree, ridge, and linear regression algorithms in distinct research work. The output from each of the aforementioned algorithms has been positive.

V. Matey, et al. [4] used a graphical depiction of price discrepancies between real data positions and a suggested best model. For the best outcome, the data was standardized. For prediction, the Linear Regression technique was utilized in this research. They described how linear regression works. This paper's technique is intriguing, and the authors clarified the relationship between the independent and dependent variables. Additionally, they discussed the RSS (Residual Sum of Squares) approach and other indicators. Using linear regression, this study obtained a minimal prediction[4] error of 0.3713. They provided a visual explanation of pricing variations from the best-fit line. Using decision trees and random forest regression, Rushab Sawant, et al. [2] forecasted home prices and provided an explanation. For estimating the property's pricing value, Z. Li [6] used four regression methods, including Support Vector Machine (SVM), Random Forest Regression and Linear Regression as well as an ensemble approach by merging SVM and Random Forest models. Support Vector Regression [9] can also be applied to data for getting the better accuracy in prediction of house values.

III. PROJECT JUSTIFICATION

The subject of real estate and housing markets places a great deal of emphasis on the problem of home price prediction utilizing machine learning techniques. For a variety of stakeholders, including purchasers, sellers, real estate brokers, and investors, accurate house price forecasts can offer insightful information. Homebuyers may use it to make well-informed judgments about paying a reasonable price for a property, sellers to set fair asking prices, and investors to gauge the success of real estate investments.

In the aforementioned studies, the issue of predicting home values using regression approaches, neural networks, support vector regression, random forest models, and XGBoost algorithms [8] is discussed. The objective of this research is to create models that can accurately

evaluate a collection of characteristics or factors related to a property, such as location, size, number of rooms, amenities, and neighborhood characteristics, and forecast the corresponding house values.

In order to create precise predictions, this research employs machine learning models to recognize intricate patterns and connections in the information. The goal variable is the appropriate sale price, and the models are trained using historical data in which the features are the traits of properties that have already been sold. Using the attributes of new or impending properties as a basis, the models may be used to forecast their pricing once they have been trained.

There are several uses for precise house price prediction models in real life. They may help prospective homeowners determine the fair market worth of a home, which enables them to negotiate pricing and make wise selections about which houses to buy. These models can help real estate agents deliver correct price advice to their consumers and enhance the overall quality of their services. These models may also be used by investors to assess the viability and possible return on investment of real estate developments.

Overall, the creation of reliable machine learning models for predicting house prices is essential for enabling informed choices in the real estate market. In order to address the issue and increase the precision of home price projections, the articles cited in the question make contributions to this field by proposing various methodologies and algorithms.

IV. TECHNICAL ASPECTS

A. SYSTEM ARCHITECTURE

Methods describe all of the essential procedures necessary to get the desired result. As seen in Figure 1, there are different phases. Used the Bangalore Kaggle dataset and preprocessed it to remove unnecessary data and characteristics. Before using algorithms, the data should be analyzed once more. And for the evaluation, used multiple regression models such as XG Boost, Random Forest Regression and, SVR (Support Vector Regression).

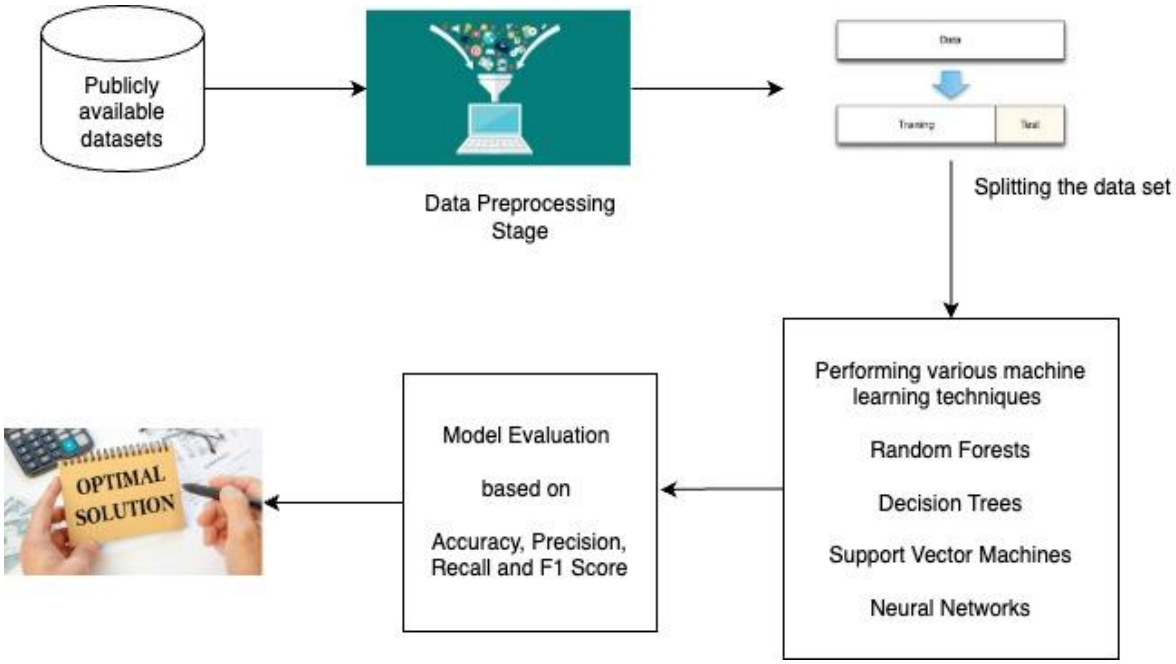

Figure 1: System Architecture

Phase 1 - Dataset

For our study, used a dataset (Table 1) from Kaggle to estimate the price of house in metropolitan cities [15]. It's a big dataset with 13330 records and 8 features includes house price, type of the area, house availability, location of the house, Number of halls, kitchen come under size, total square feet of the house , bathroom, and society are among the qualities.

Features	Description	Data Type
area_type	Type of area	Non-numerical
availability	Date of availability	Non-numerical
location	Address of house	Non-numerical
size	No. of room, hall, and kitchen	Non-numerical
society	Society	Non-numerical
total_sqft	Total area in Square feet	Non-numerical
bath	No. of Bathroom	Numerical
balcony	No. of Balcony	Numerical
price	Price of house	Numerical

Table 1: Features and dataset types

Phase 2 - Data Cleaning, preprocessing, and training the data

Data cleansing is the process of removing unneeded and incorrect data from a gathered dataset. Various tools may be used to clean up data and remove such junk values. It discovers such values and substitutes the jumbled data. This data is changed to guarantee that it is correct and precise. The primary goal of cleaning the data is to identify and eliminate incorrect values in order to develop dynamic information estimates. The cleansed data must then be put in a fresh dataset. As a result, after collection of the data, data cleansing is a vital step to do in order to obtain reliable findings.

This approach comprises dataset preprocessing and splitting the data into train and test data. Can see non numerical characteristics in the dataset such as location of the house, condition of the house, see the air transmission in room, and so on. Used the sci-kit learn library's One Hot Ender and label Encoder functions to transform these non-numerical properties to numerical ones. Also removed less related information like type of the area. Dataset contains the empty cells; replaced those cells with the column mean by applying the Simple Imputer function which is the part of the sci-kit learn module.Selling price of the house is the target feature in this dataset. Next step is to split the data into train and tests, data can be split using the test_split function. Data can be splitted based on the 8:2 ratio means 80% train data and 20% test data.

At this point, the cleaned data is used to generate test and train data. The data that is trained is the bigger dataset that was utilized for training the model's algorithms. The desired value will be included in the training set as well.

Before passing the data to any model, must ensure entire information is correct and set for use. For this, assessed our information using these features and the relationship between the features.

Phase 3 - Model Validation

The procedure for the validation contains decides to check whether that particular model is suitable for that particular dataset. The fundamental motivation is to obtain the maximum level of accuracy possible. Following the initial method, more algorithms can be tested on that dataset to evaluate which generates the best accuracy. Input-output data is used to represent the model. The validation test compares whether the input parameters matches with the output data that is generated when a particular model is applied to a dataset. Accordingly, the results are recorded.

B. DESCRIPTION OF MODELS

1. Linear Regression

Importing the libraries and dependencies, such as sklearn, which provide tools for doing linear regression. Import the linear regression module from the selected library [3]. Set the labels for the regression model to the target variable, which in this case is the home price column. Any extra dependencies or modules necessary for preprocessing or data manipulation should be imported. Divide the dataset into train and test data. This is necessary to assess the model's performance using previously unknown data. Dataset can be divided into two parts: features (input variables) and labels (house prices). Train the linear regression model using the train data. The model recognizes the link between the properties and the target variable. Train the linear regression model using the train data. The model recognizes the link between the properties and the target variable. Once the model has been trained, utilize the testing data to predict housing prices. The model estimates house prices by applying the obtained coefficients to the features.

2. Decision Tree

Decision Tree comes under the category of supervised machine learning algorithm [4]. Decision tree can estimate the price of a property with high accuracy. In this approach, will keep the model in the form of a tree, which will be useful for all future predictions. It learns by

analyzing system data in the form of maximum and minimum values. A decision tree is a data structure that is comparable to a tree data structure that contains 4 parts.

a. The decision tree's root node is the Internal Node.
b. A test that is conducted on feature values is indicated.
c. A branch indicates the rules of the decision tree.
d. Node LeafThe conclusion of the decision tree.

Entropy and the Ginni index provide the decision tree's mathematical core.
Entropy is stated as the size of information needed to precisely characterize specific samples. If Entropy is 0, then it indicates that there are similar samples; entropy is one if there are different samples.

Advantages of the decision tree:

a. Helps in the simplification of the information.
b. The decision tree gives a stronger idea of which information is relevant for predicting.
c. Nonlinear information has no effect on model performance.

3. XG Boost

XG Boost [5] belongs to the group of supervised machine learning algorithms. This model helps in merging the weak learners to generate the strong learners for improving the accuracy of the models. XG Boost is an advanced model and it is a gradient boosting method which is specifically designed to work on large and complicated datasets. For reducing the performance time and to get the least time and blend the optimization of software and hardware methodologies for getting the better results. Initial learners are produced successively in this specific type of boosting, current starting learners are always more successful than the preceding ones.
The XG boost consists of 3 primary parts:

a. Improvement is required for the loss functions
b. Untrained in prediction calculation and formation
c. XG Boost is represented as a additive models with Normalized Loss Functions.

The XG boost supervised Machine Learning Model l is the combination of numerous weak learners. W=w1,w2,w3,.....wn collection of inefficient learners

4. Support Vector Regression

SVR [9] is a type of supervised machine learning algorithm. Suppose, lets consider if I have 'm" columns, plot them in an m-dimensional plane, with each column's value indicating a specific coordinate.

The primary goal of SVR is to find the best hyperplane (point in one dimension, line in 2 dimensions, plane in 3 dimensions, and hyperplane in n dimensions) for linearly separating data points into two components while the increasing the margin.

SVR has the following advantages:

a. It is successful in larger size when the number of samples is less than the number of magnitudes.
b. Make use of kernels such as linear kernels and Radial basis kernels (RBG).
c. Model performs best with a distinct dividing margin.
d. Primarily used in classifying the image and bioinformatics, text and hypertext categorization, and face identification.

5. Random Forest

The random forest [7] is an example of supervised machine learning algorithms. It functions as a starting method with which combine multiple basic learners to create a powerful model. In the case of random forest, base learners are the decision trees, and take an average of their continuous values.

Lets see how the decision tree works.

If r is less than three, R(q1) = 10.

1. If r is less than 6 but larger than 3 decision tree,
R(r2) =20. 2. If r is bigger than 6,
R(r3) = 50 decision tree will be 3.
The random forest is defined as
Y = k R(r1) + R(r2) + R(r3). (4) Where 'k' denotes a constant.

Advantages

a. Performs greater on huge and complex datasets.
b. This model is described as the more accurate machine learning algorithms.
c. It operates on a large number of variables without omitting any of them.

6. Lasso Regression

Import the libraries and dependencies required for implementing Lasso regression [2], such as sklearn. Import the Lasso regression module from the selected library.Set the labels for the regression model to the target variable, which is the home price column.Import any

additional dependencies or modules necessary for preprocessing or data manipulation should be imported.
For evaluating the performance of the unknown data, dataset is splitted into train and test data. Dataset is divided into two parts features (input variables) and labels (house prices).
Train the Lasso regression model using the trained data. The model sees the link in between the features and the target variable while using L1 regularization to reduce the coefficients of minor features to zero.Once the model has been trained, utilize the testing data to forecast housing prices. The model estimates house prices by applying the learnt coefficients to the characteristics while accounting for the regularization term. The Lasso regression procedure comes to an end, and the anticipated housing values may be reviewed for accuracy and future investigation.

When working with datasets with a high number of features or when feature selection is necessary, Lasso regression is especially beneficial. Lasso regression supports sparsity by including a penalty component in the loss function, which means it prefers to choose the most significant features while setting the coefficients of less essential characteristics to zero. This can aid in determining the primary elements that influence property values.

C. Implementation Plan

The following are the steps in the implementation plan for house price prediction using machine learning techniques. To begin, gather the dataset to train and test the models, which can be gathered from publicly available sources or real estate databases. Location, number of rooms, area, facilities, and historical pricing data should all be included in the dataset. After collecting the dataset, it must be preprocessed by dealing with missing values, outliers, and conducting feature engineering, such as one-hot encoding categorical variables and scaling numerical features.

Following that, the dataset must be separated into training and testing sets, often in the 70:30 or 80:20 ratio. This ensures that the models are trained on enough data and assessed on unknown data to appropriately estimate their performance.

Various regression techniques, including as linear regression, support vector regression (SVR), random forest, XGBoost, and neural networks, can be used in the implementation. Each algorithm must be developed and tested on the training dataset. Popular machine learning libraries such as scikit-learn, TensorFlow, and Keras may be used to create the models. To discover the ideal hyperparameters for each model, hyperparameter tuning approaches such as grid search or randomized search may be used throughout the training phase.

After the models have been trained, they must be assessed using appropriate metrics such as mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), or

R-squared score. To choose the most accurate and dependable model for house price prediction, the models should be evaluated based on their performance.

Once the best model has been chosen, it must be deployed to make predictions on previously unknown data. To be utilized in a production setting, the model can be saved or serialized. A user interface may be created using web technologies or a mobile application to allow users to enter important house attributes and receive a forecasted price based on the trained model. The deployed model should be updated and retrained on a regular basis with new information to make sure its correctness and relevance over time.

D. Limitations of Models

The SVR (Support Vector Regression) model, which is used to predict home prices, has the feature of being sensitive to the kernel function and its hyperparameters, which can affect its predictive ability. Linear regression approaches presume a linear connection between the input characteristics and the target variable, which may be insufficient for capturing complicated nonlinear patterns in home price data. Depending on the depth and complexity of the tree, decision tree models can suffer from overfitting or underfitting, and they can be sensitive to slight changes in the training data. While strong and efficient, XGBoost models may require careful hyperparameter tuning and may be prone to overfitting if not accurately regularized, which may compromise their predictive performance for home price prediction applications.

V. CONCLUSION AND FUTURE SCOPE

Customer satisfaction is boosted through improving in making the decision accuracy and reducing the dangers associated in purchasing the real estate properties. A more precise and accurate sales price will be employed. The technology can delight clients by providing precise findings and decreasing the danger of making the incorrect investment. Using the machine learning in the real estate industry is still in the early stages. Hope that the models had made a little contribution to the property field appraisal by giving some methodological and empirical contributions, as well as an alternate way for calculating housing expenditures. According to my study, XG Boost has the highest accuracy in forecasting property prices based on the metrics of root mean squared error (RMSE), mean absolute error (MAE), and mean squared error (MSE) followed by decision tree and linear regression.

To begin, most research concentrate on particular regions or localities, that might restrict the appropriateness of their results to other areas. When applied to diverse places, the local features and dynamics of housing markets can have a substantial influence on the prediction models' success. Data quality and availability might be problematic. It can be difficult and time-consuming to get reliable and complete datasets that include essential elements such as

property attributes, economic indicators, and neighborhood information. Inadequate or inadequate data may result in unbalanced or inaccurate forecasts. Machine learning algorithms and regression techniques used may have inherent limitations. These models' success is largely dependent on good feature selection, parameter adjustment, and management of outliers or missing information.

It is recommended that future studies look into complex ensemble methods which include different models to improve prediction performance. Incorporating more diverse and wide datasets, such as socioeconomic characteristics, geographical information, and market trends, can also contribute to more accurate forecasts. Deep learning technologies such as convolutional and recurrent neural networks might be investigated to capture complicated patterns in housing data. It would also be useful to evaluate the performance of these models using real-world datasets from various areas and marketplaces.

REFERENCES

[1] A. P. Singh, K. Rastogi and S. Rajpoot, "House Price Prediction Using Machine Learning," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 203-206, doi: 10.1109/ICAC3N53548.2021.9725552.

[2] V. S. Rana, J. Mondal, A. Sharma and I. Kashyap, "House Price Prediction Using Optimal Regression Techniques," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2020, pp. 203-208, doi: 10.1109/ICACCCN51052.2020.9362864.

[3] J. Manasa, R. Gupta and N. S. Narahari, "Machine Learning based Predicting House Prices using Regression Techniques," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 624-630, doi: 10.1109/ICIMIA48430.2020.9074952.

[4] V. Matey, N. Chauhan, A. Mahale, V. Bhistannavar and A. Shitole, "Real Estate Price Prediction using Supervised Learning," 2022 IEEE Pune Section International Conference (PuneCon), Pune, India, 2022, pp. 1-5, doi: 10.1109/PuneCon55413.2022.10014818.

[5] C. Sheng and H. Yu, "An optimized prediction algorithm based on XGBoost," 2022 International Conference on Networking and Network Applications (NaNA), Urumqi, China, 2022, pp. 1-6, doi: 10.1109/NaNA56854.2022.00082.

[6] Z. Li, "Prediction of House Price Index Based on Machine Learning Methods," 2021 2nd International Conference on Computing and Data Science (CDS), Stanford, CA, USA, 2021, pp. 472-476, doi: 10.1109/CDS52072.2021.00087.

[7] Y. Fang, T. Li and H. Zhao, "Random Forest Model for the House Price Forecasting," 2022 14th International Conference on Computer Research and Development (ICCRD), Shenzhen, China, 2022, pp. 140-143, doi: 10.1109/ICCRD54409.2022.9730190.

[8] C. Chee Kin, Z. Arabee Bin Abdul Salam and K. Batcha Nowshath, "Machine Learning based House Price Prediction Model," 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 2022, pp. 1423-1426, doi: 10.1109/ICECAA55415.2022.9936336.

[9] Z. Yi, Z. Chunguang, H. Lan, W. Yan and Y. Bin, "Support Vector Regression for Prediction of Housing Values," 2009 International Conference on Computational Intelligence and Security, Beijing, China, 2009, pp. 61-65, doi: 10.1109/CIS.2009.127.

[10] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.