House Price Prediction Review

Smit Chatwani
Information Technology
Dwarkadas J. Sanghvi College of Engineering Mumbai, India

Mayank Chotaliya
Information Technology
Dwarkadas J. Sanghvi College of Engineering Mumbai, India

Vinaya Sawant
Information Technology
Dwarkadas J. Sanghvi College of Engineering Mumbai, India

ABSTRACT
The property market is among the most price-sensitive in the world, and it is constantly shifting. It is one of the most essential topics where machine learning techniques could be used to predict better and accurately. Physical conditions, concepts, and location are the three variables that impact the price of a home. The existing approach includes evaluating house values without considering market pricing or cost inflation. Future prices will be reduced by analyzing past market patterns and value ranges, as well as future improvements. This paper includes the study of various prediction models**.**

Keywords
Real Estate Management, Price Prediction.

1. INTRODUCTION
Many real estate organizations focus more on a project to gain from the growing competition in the real estate sector. Improve the decision-making process for projects, the foundation is the execution and functioning of several stages of excellent management to achieve the goal of speedy. The project's development is desired, as is the project's correction. Management framework, achieve management optimization system. The Residential Sector accounts for the majority of the real estate market. The country's real estate industry. Inflationary pressures, rising incomes. An increase in the number of white-collar workers because of urbanization. Professionals, for example, are the primary motivating elements. This segment is to blame for the company's growth.

Luxury and super-premium residences are once again in high demand among globe-trotting executives, new and successful businessmen, non-resident Indians (NRIs), and others in the residential market. With the rise in real estate demand comes the issue of finishing projects on time. As a result, the action plan must be analyzed because there are numerous processes involved in project management that are dependent on various aspects such as the owner, the contractor, the project's risk, the decision-making process, and so on.Technical articles published in proceedings and other journals are mentioned to identify the scope of the task and to keep track of the progress of each project. It should also be noted that many academics and researchers have worked on the real estate development process.

Real estate is one of the most price-conscious industries in the world, and it is always evolving. It's one of the most crucial areas where machine learning methods may help enhance and precisely anticipate expenses. Physical conditions, concepts, and location are all elements that determinea house's price. The existing framework comprises assessing house values without taking into account market prices or cost inflation. The paper's goal is to forecast residential pricing for clients based on their financial objectives and requirements. Future expenses will be reduced by analyzing past market patterns and value ranges, as well as future technological improvements.

Machine learning tries to create self-learning algorithms based on datasets so that machines can forecast future behavior based on historical data. It aids organizations in quickly recognizing and forecasting trends and patterns. It also enables managers, executives, and analysts to use its models to make more efficient decisions. It enables the organization to make environmental changes without the need for human intervention.

Machine learning is an algorithm that is used to create a model and then use that model to predict fresh data sets. The main difference between using a normal algorithm and utilizing one that is oriented with the input data rather than focused on a chain of various instruction sets is that the result is oriented with the input data. Unsupervised learning is entirely focused on unlabeled data sets, whereas supervised learning is entirely focused on constructing a model based on labeled data sets. Regression, classification, clustering, SVM, neural networks, deep learning, and other machine learning algorithms are examples. It is critical to anticipate a certain outcome using a feature extraction-based model. As far as home prediction is concerned, the following metrics are used to determine it:

- location,
- dimension,
- house type,
- city,
- country,
- tax rules,
- economic cycle,
- population movement,
- Interest rate, and a lot of other elements which could impact demand and supply.

With the existing linkages with so many known and unknown features, the real estate sector is subject to wide price volatility all over the world. These property values can increase or drop at variable rates throughout time, depending on the market and non-market concerns. The goal of this project is to create a machine learning application that would identify and improve price projections for housing units, allowing buyers to make worthwhile real estate investments.

After agriculture, the housing sector in India is the second-largest employer, with a degree of urbanization of 33.54 percent (Statista, 2018), and is predicted to contribute 13% of the country's GDP by 2025. Every year, 10 million people relocate to cities, with young people (15–35 years old) accounting for 35% of the population. Because ambiguity in

house prices makes it difficult for buyers to choose their dream home, a definitive housing price prediction model can help buyers, sellers, and real estate agents make better-informed decisions.

In this research, an attempt was made to anticipate decisive house pricesin India, so that potential purchasers might make purchasing decisions based on the projections. Various machine learning tools are used to select an appropriate prediction method. To identify efficient price prediction models, a variety of machine learning methods were evaluated, including regression, decision trees, k-nearest neighbors, random forest, and support vector machines, among others.

2. RELATED WORKS

House prices are influenced by several things. Rahadi, et al. [1], in their study, separate these characteristics into three categories: physical condition, concept, and location.
The size of a house, the availability of a kitchen and garage, the number of bedrooms, the area of land and buildings, the availability of a garden, and the age of the house are all physical conditions that can be observed by human senses [2],A concept, on the other hand, is a concept offered bydevelopers to entice potential buyers, such as a home with a pool.

The hedonic price theory states that a property's value is equal to the sum of its attributes' values [3]. Hedonic pricing is a price prediction model based on the hedonic price theory. A regression model can be used to achieve hedonic pricing. Equation 1 depicts the regression model for determining a price.

$$y=a.x_1 +b.x_2 + \ldots + n.x_i \quad (1)$$

Where y represents the predicted price and x1, x2, and xi represent the characteristics of a home. The correlation coefficients of each variable in the determination of housing prices are denoted by a, b,... n.

Raghunandan [4] discussed the fundamentals of data mining, including how it works and supporting algorithms for prediction. The most crucial aspect is determining which machine learning algorithm is appropriate for predicting house prices. Manjula [5] highlights numerous significant features to employ when projecting property prices with good precision using a regression model because the location's environmental conditions often determine what kind of price we can expect for different types of residences. A. Varma [6] devised a method for obtaining precise real-world appraisals using Google maps and real-time local data.

Researchers also discovered that there are links between physical appearance and non-visual characteristics such as statistics on crime, housing prices, population density, etc. For example, “City Forensics: An Introduction” is a book about forensics in cities predicting Non-Visual City Using Visual Elements Visual attributes are used to predict the sale in “Attributes” [7].Hujia Yu and Jiafu Wu (2014) [8] used classification and regression algorithms. The living space square feet, roof content, and neighborhood have the most statistical importance in predicting a home's selling price, according to the data. Prediction analysis can also benefit from the PCA technique. Li Li and Kai-Hsuan Chu (2017) investigated backpropagation neural networks (BPN) and radial basis functional (RBF) neural networks [9]. RBF and BPN models are utilized to recognize the difference between house price indexes such as Cathy and Sinny price indexes and sophisticated correlation functions to detect macroeconomic analysis.

3. LITERATURE REVIEW

Technology has completely transformed the construction industry throughout the years. The location (POS) operations have seen a lot of innovation. Several definitions are suited for economic feasibility assessments that are derived from the applied economy's contents. The applied economy derives its philosophy and methodology from economic science and political economy theory, which integrates various sciences such as accounting, business administration, knowledge systems, law and research, to achieve investment rationalization based on the feasibility of an economic project

Some economists use the feasibility study as a technique for investment decision-making that relies on a set of tactics, tools, tests, and scientific foundations that rely on accurate information about a project's possible failure. The feasibility study evaluates the project's ability to achieve goals focused on maximizing revenue and profit for the investor's non-public and financial systems, or both, over time.Some economists define it as a collection of studies aimed at determining the viability of an investment project or a collection of investments that include a wide range of market, technical, financial, economic, and social elements to choose the best approach. Others define it as complete scientific investigations of all components of a project or expected outcome, which can be found within a variety of extended preliminary studies that lead to a decision or investment opportunity among numerous options or projected investment chances. This research should be accurate, objective, and thorough. They are a collection of specialized studies scattered to verify that project outcomes (benefits and profits) exceed or are appropriate in relation to their inputs (costs). Many have defined the project's fallibility as research that represents the potency or sufficiency of a projected investment that is assessed on an analytical basis of available options to make the best decision. This selection will be based on the use of factors or monetary income and price measures, as well as the time required to fulfill initial obligations or the inherent loss of essential values. Furthermore, it should be a strictly industrial or national economic analysis by the decision-making principles that are also linked to the standards of the proposed project or its strategic planned statement within the country.

Regression analysis is a predictive modeling approach that examines the relationship between a dataset's goal or dependent variable and its independent factors. It entails finding the best-fitting line that passes through all of the data points with the least amount of gap between the line and each data point. We are experimenting with several regression techniques on a specific issue statement to discover the best fitting model for the most accurate predictions. Linear regression, Support Vector Regressors and Decision Trees are examples of this.

Linear Regression- The basic goal of the Linear Regression model is to determine the best-fit linear line as well as the appropriate intercept and coefficient values to minimize error. The discrepancy between the actual value and the predicted value is defined as an error. The goal is to lessen the difference or inaccuracy. Simple and Multiple Linear Regression are the two variants of linear regression based on the number of independent variables. In Simple Linear Regression, there is only one independent variable, and the model must find a linear relationship between it and the

dependent variable. Many Linear Regression, on the other hand, looks for a link between the dependent and independent variables by using multiple independent variables.

We predict scores on one variable from scores on the second variable in basic linear regression. The variable we're seeking to forecast is the criteria variable, sometimes known as Y. The predictor variable, abbreviated as X, is the variable on which our predictions are based. When there is only one predictor variable, simple regression is utilized to make predictions. When the predictions of Y are plotted as a function of X in simple linear regression, which is the topic of this section, they form a straight line.

Figure 1 depicts the example data from Table 1. There is a positive association between X and Y, as you can see. If you were to forecast Y based on X, the greater the value of X, the more accurate your prediction of Y would be.

Table 1. Example data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

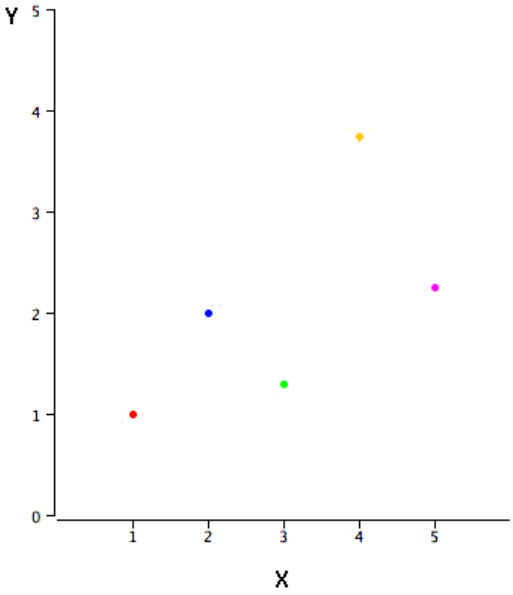

Figure 1. A scatter plot of the example data.

Finding the best-fitting straight line across the points is the goal of linear regression. A regression line is a best-fitting line. The regression line in Figure 2 is the black diagonal line that contains the expected Y score for each possible value of X. The prediction errors are represented by the vertical lines connecting the points to the regression line. The red point, as you can see, is extremely close to the regression line; its prediction error is modest. The yellow point, on the other hand, is substantially higher than the regression line and hence has a large prediction error.

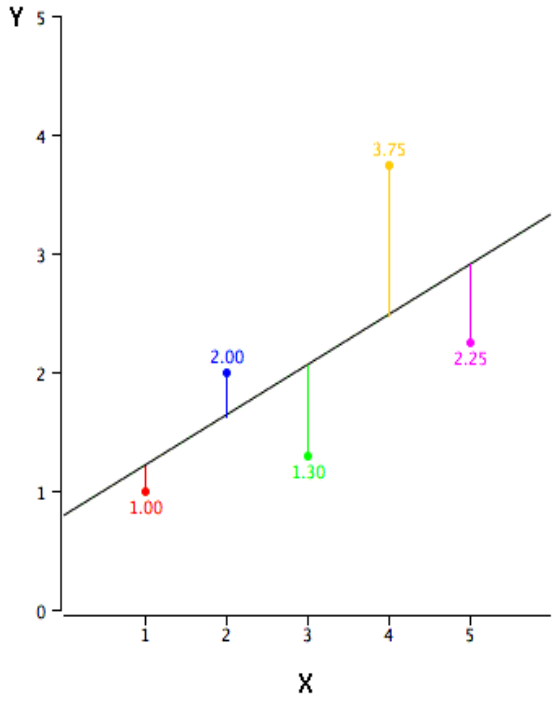

Figure 2 shows a scatter plot of the data in the example. The predictions are represented by the black line, the dots by the actual data, and the vertical lines between the points and the black line by the prediction errors.

The value of the point minus the projected value is the prediction error for a point (the value on the line). The prediction errors (Y-Y') as well as the predicted values (Y') is displayed in Table 2. The first point, for example, has a Y of 1.00 and a forecasted Y (referred to as Y') of 1.21. As a result, its prediction error is -0.21.

SVMs (Support Vector Machines) are a group of supervised learning algorithms for regression, classification and outlier detection. These are all common machine learning tasks. They can be used to detect cancerous cells using millions of images or to forecast future travel patterns using a well-fitted regression model. Support vector regression (SVR), which is an extension of support vector classification, is one sort of SVM you can employ for certain machine learning problems (SVC). The most important thing to remember is that they are just arithmetic equations that have been fine-tuned to provide you with the most accurate answer possible as rapidly as feasible. SVMs are distinguished from other classification algorithms by their decision boundary, which maximizes the distance between all classes' nearest data points. The decision boundary established by SVMs is known as the maximum margin classifier or maximum margin hyperplane.

SVMs are employed in machine learning for a variety of reasons. Handwriting recognition, intrusion detection, face identification, email categorization, genre classification, and web page classification are all applications that use SVMs. This is one of the reasons why we use SVMs in machine learning. It can perform classification and regression on both

linear and non-linear data.SVMs are also used because they can detect complex relationships between your data without requiring a lot of user input. When working with smaller datasets with tens to hundreds of thousands of features, it's a wonderful solution. Because of their ability to handle small, complex datasets, they usually produce more accurate results than other algorithms.

Here are some of the advantages and disadvantages of utilizing SVMs.

Pros-

Effective on datasets having a variety of characteristics, such as financial or medical data.

When the number of features exceeds the number of data points, this method is effective.

Support vectors are a subset of training points used in the decision function, which makes it memory efficient.

For the decision function, different kernel functions can be chosen. You can use standard kernels, but you can also provide custom kernels.

Cons-

When the number of features exceeds the number of data points, it's critical to avoid over-fitting when selecting kernel functions and regularisation terms.

Probability estimations aren't directly provided by SVMs. These are computed using a time-consuming five-fold cross-validation method.

Decision Tree- A Decision Tree is a supervised learning method that can be used to solve classification and regression problems, but it is most commonly used to solve classification issues. In this tree-structured classifier, internal nodes hold dataset attributes, branches hold decision rules, and each leaf node holds the conclusion. A Decision Tree has two nodes: a Decision Node and a Leaf Node. Leaf nodes are the outcome of such decisions and have no additional branches, whereas Decision nodes are used to make any decision and have many branches. To make judgments or run tests, the features of the given dataset are used.

It's a visual depiction of all the possible solutions to a problem or decision based on a set of criteria. It's called a decision tree because it begins with a root node and grows into a tree-like structure with many branches, much like a tree. To create a tree, we use the CART algorithm (Classification and Regression Tree). A decision tree is a tree that asks a question and splits into subtrees based on the answer (Yes/No).

Decision Tree Terminologies- The decision tree's root node serves as its beginning point. It's used to represent the whole dataset, which is then split into two or more homogeneous groups. Leaf Node: The tree's final output node that cannot be split. Splitting is the method of dividing the decision node into sub-nodes based on the provided conditions.

A tree that has been split into branches or subtrees.

Pruning is the procedure of pruning a tree to remove undesired branches. The root node of the tree is known as the parent node, while the remaining nodes are known as the child nodes.

4. PROPOSED SOLUTION

E-learning and e-education are now extensively impacted. Manual processes are being phased out in favor of automated ones. This project's purpose is to forecast house prices to minimize the customer's problems. The current strategy entails the customer approaching a real estate agent to handle his or her investments and recommend suitable estates. However, this strategy is hazardous because the agent may estimate incorrect estates, resulting in the customers' capital being lost. The manual approach now in use on the market is obsolete and fraught with danger. There is a need for an updated and automated system to address this flaw. Data mining algorithms can be used to assist investors in selecting an acceptable property based on their stated needs. The new system will also save money and time. The operations will be simple. The suggested system is based on the nave Bayes classification algorithm. The system will estimate the house price depending on the information entered by the administrator. When a user searches for a property, a list of properties with estimated prices is provided. The user may sell his house by putting his details into the system, and he can also utilize our suggested method to find a rental home.

5. DATASETS

In this paper, we used two datasets and applied multiple known machine learning algorithms to them to predict prices.

A.The first dataset, which covers housing values in Boston suburbs, comes from the UCI Machine Learning Repository. This data was derived from Carnegie Mellon University's StatLib library. The label/price is predicted using attribute variables because this article uses machine learning for price prediction. The collection of attribute variables used to create the prediction model is shown in the table below. For predicting property values, this study includes 13 qualities as independent variables.

Table 1: Attributes and labels in the dataset (Boston)

Attributes	Description
CRIM	per capita crime rate by town
ZN	The proportion of residential land zoned for lots
INDUS	The proportion of non-retail business acres per town
CHAS	Charles River dummy variable
NOX	nitric oxides concentration (parts per 10 million)
RM	The average number of rooms per dwelling
AGE	The proportion of owner-occupied units built before 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63) ^2 where Bk is the proportion of blacks by

	town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $ 1000's

B. Anthony Pino gathered this data from publicly available results posted every week on Domain.com.au.

For the price forecast, we used 13 important influencing factors.

Table 2: Attributes and labels in the dataset(Melbourne)

Attributes	Description
Council Area	Governing council for the area
Method	Method of sale
Region name	General region
Rooms	Number of rooms
Type	Type of house
Distance	Distance from CBD in Kilometres
Bedroom2	Scraped # of Bedrooms (from different source)
Bathroom	Number of Bathrooms
Car	Number of carspots
Land size	Land Size in Metres
Building Area	Building Size in Metres
Year Built	Year the house was built
Property count	Number of properties that exist in t suburb
Price	Price in Australian dollars

6. CONCLUSION

The desire to buy a home is shared by everyone. We want people to be able to buy houses and real estate at their true value utilizing this recommended approach, rather than being tricked by dodgy agents looking for a quick buck.

Furthermore, this approach will benefit huge organizations by offering precise estimates that will help them to establish prices while saving them time and money. The essence of the market is accurate real estate values, which we aim to achieve through this strategy.

Using the raw data supplied to it, the system is capable of self-training and price prediction. After you've finished reading,a collection of research papers, blogs, and articles,There were several algorithms chosen that were suitable for useon both of the model's datasets Following extensive testing and evaluation,After several training sessions, it was discovered that the Linear Regression Among the other algorithms, produced the best results. The system was capable of foreseeing the future.I was able to compare the pricing of numerous houses with varying characteristics and was ableto deal with massive amounts of data The technology is really easy to useas well as saves time.

The additional feature that can be added to our suggested system is to provide users with a full-featured user interface, allowing them to use the ML model for many locations with multiple functionalities. Additionally, an Amazon EC2 connection will enhance the system's functionality and convenience of use. Finally, the project will be completed by creating a well-integrated online application that can anticipate prices whenever users want it to.

7. REFERENCES

[1] 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.

[2] V. Limsombunchai, ―House price prediction: Hedonic price model vs. artificial neural network,‖ Am. J. …, 2004.

[3] S. Rosen, ―Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition,‖ J. Polit. Econ., vol. 82, no. 1, pp. 34–55, 1974.

[4] Lakshmi, B. N., and G. H. Raghunandhan. "A conceptual overview of data mining." 2011 National Conference on Innovations in Emerging Technology. IEEE, 2011.

[5] Manjula, R., et al. "Real estate value prediction using multivariate regression models." Materials Science and Engineering Conference Series. Vol. 263. No. 4. 2017.

[6] A. Varma et al., "House Price Prediction Using Machine Learning And Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies, pp. 1936–1939, 1936.

[7] Arietta, Sean M., et al. "City forensics: Using visual elements to predict non-visual city attributes." IEEE transactions on visualization and computer graphics 20.12 (2014): 2624-2633.

[8] Yu, H., and J. Wu. "Real estate price prediction with regression and classification CS 229 Autumn 2016 Project Final Report 1–5." (2016).

[9] Li, Li, and Kai-Hsuan Chu. "Prediction of real estate price variation based on economic parameters." 2017 International Conference on Applied System Innovation (ICASI). IEEE, 2017.