International Journal of Innovative Research in Engineering
Volume 4, Issue 3(May-June 2023), PP: 461-466.
www.theijire.com **ISSN No: 2582-8746**

Predicting Real Estate Price Using Linear Regression

Avinash Singh[1], Vinayak[2], Rudrendra Bahadur Singh[3], Anshuman Yadav[4]
[1,4]Students, Babu Banarasi Das Institute of Technology and Management, Lucknow, Uttar Pradesh, India.
[2,3]Assistant Professor, Babu Banarasi Das Institute of Technology and Management, Lucknow, Uttar Pradesh, India.

How to cite this paper:
Avinash Singh[1], Vinayak[2], Rudrendra Bahadur Singh[3], Anshuman Yadav[4], "Predicting Real Estate Price Using Linear Regression", IJIRE-V4I03-461-466.

Abstract: *The purpose of the paper is to predict the market value of the property being sold. This program helps to find the starting price of a location based on location variables. Similarly, consider a situation in which a person needs to sell a house. By using a real estate pricing system, the seller will be able to determine what features he can add to the house so that the house can be sold at a higher price. Therefore, in both cases, we can be sure that the home price is good for both the buyer and the seller. Housing prices go up every year, so there is a need for a real estate forecasting system. Estimating the price of a house can help a developer determine the selling price of a house and can help clients set a reasonable time to buy a home. Buying a house is one of the biggest financial goal of everyone. Owning a house is not only a basic need but it also represents prestige. However, buying a house is one of the most crucial decision of a person's life as there are so many factors to be consider before buying a property. House prices keeps changing based on location, area, population, house condition and structure, availability of parking, backyard, size of house etc. From past few years a lot of data has been generated regarding Real Estate. Machine learning prediction techniques can be very useful to predict an accurate pricing of the houses. The study focuses on developing an accurate prediction model for house price prediction. Machine learning is sub-branch of artificial intelligence that deals with statistical methods, algorithms. Using machine learning we can build a model which can make prediction based on past data. In this paper we will review different machine learning algorithms which can be used for house pricing prediction.*

Key Word: *Machine learning models, house price prediction, real estate, price prediction, Machine learning algorithms.*

I.INTRODUCTION

Nowadays, many people invest in real estate, because sometimes, it can bring a lot of capital income, which has developed very violently in our country. Investment in real estate is also a reflection of a local real estate development situation to a certain extent. For example, if investment is hot, it means that the development is unbalanced, and the supply is in short supply. If investment is cold, it means that the recent real estate market is relatively stable. Moreover, it is also a reflection of the development of a city. However, it does not mean that the more the investment, the better the real estate will develop, the more benefits the investors will get, and there may be negative situations, and some places have suffered from this situation. Moreover, real estate investment in many places has gone wrong, exceeding demand. This is a waste of resources and irresponsible for people's lives, and there will also be situations where workers are not paid, causing social chaos. Therefore, in order to prevent these situations from happening, it is very important to control the investment in real estate. In the process of urbanization, the government should actively formulate reasonable measures to control the situation and ensure that the proportion of investment in fixed assets remains around 25%. Moreover, this limit is not fixed. It also depends on the development of the city. After the initial stage of urban development, this ratio can be appropriately reduced, because there is no need for so many houses at this time. What we need to know very clearly is that we cannot make reasonable improvement measures in a timely manner. The reason for this is that we always know the problem after the situation arises, and it has a certain lag effect. Therefore, when formulating, it is necessary to fully and comprehensively consider the development situation and changes in the relationship between supply and demand, strive to achieve standards that can meet the long-term development situation, and try to avoid the rise in housing prices due to incorrect measures. This is not only a guarantee for the stable development of society, but also a guarantee for people's lives. Only when people are stable can a country develop well.

At present, the methods of housing price forecasting can be divided into two categories. One is a multifactor analysis model based on the analysis of the influencing factors of housing prices, and the other is a single-factor analysis based on time series. In the multivariate analysis models, most of them only consider the parameters that affect housing prices, such as multiple regression models, but do not consider the nonparametric factors. The absence of some nonparametric influencing factors is likely to lead to a decrease in the accuracy of the prediction model. In the process of reviewing the literature, only one paper was found that used a partial linear model to predict the average sales price of commercial housing across the country, and the results of the paper showed that the partial linear model was better than the linear regression model in predicting housing prices. This is because the partial linear model considers both linear and nonlinear factors affecting housing prices. However, considering that there are many factors affecting housing prices, there will be a curse of dimensionality when using a partial linear model.)e additive model can eliminate the disaster of dimensionality, so it is of great practical significance to build a housing price prediction model based on the additive model. In addition, it also has important theoretical significance to establish a housing price prediction model on the basis of the additive model. In different places, housing prices are affected by local policies and special events, and the fluctuation laws of housing prices are also different.

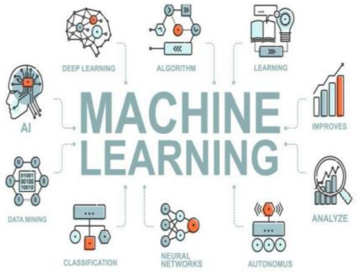
Machine Learning:

Machine learning has played a major role since years ago in selecting an image, spam, common speech command. Machine learning also provides better customer service and safer car systems. The home-market market stands out among the most focused on prices and is constantly evolving. It is one of the main areas in which to apply machine learning ideas on how to improve and predict costs with high accuracy. A single heuristic database is often used to analyse the decline in house prices in the Boston city housing database. Previous analysis has found that house prices in that database depend largely on their size and location. Until recently, basic algorithms such as linear regression could reach 0.113 predictive errors using both internal building features (living space, number of rooms, etc.) and additional spatial features (demographic image features such as revenue rating, population. congestion, etc.). Modelling uses machine learning algorithms, in which the machine learns data and uses it to predict new data.

As we know, the proposed model for accurately predicting future outcomes has economic applications, business, banking, health care sector, e- commerce, entertainment, sports etc. One such method used to calculate real estate is based on a number of factors. In large metropolitan areas such as Bengaluru, a potential buyer considers a few factors such as location, size, proximity to parks, schools, hospitals, power stations, and most importantly the value of the house.

Machine learning: Machine learning is a data-based system learning process. Machine learning is part of data science where we use machine learning algorithms to process data.

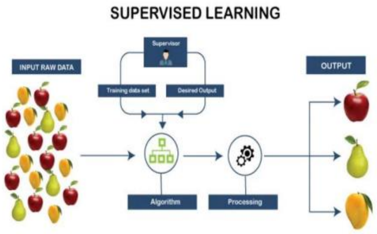
Supervised reading strategy It is a hypothetical model used for activities where it involves predicting one value using other values in a data set. Supervised reading will have pre-defined labels. Separates an object based on parameters in one of the predefined labels.



Supervised Learning Method:

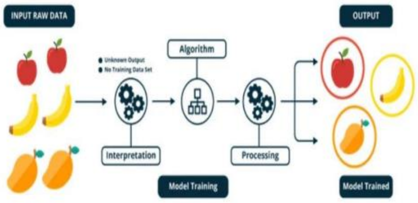
Descriptive model is used for activities that can benefit from the understanding gained by summarizing data in new and exciting ways. There are no predefined labels in unattended reading mode. In the real estate market, data is a very important source of analysis and forecasting. It is always beneficial to know about the speculation of a business diversity that will occur soon and business executives can do the right thing to avoid future losses. And because of this we need a very accurate prediction Model for analysis. Similarly, we need accurate forecasting of housing and real estate in the real estate market to provide accurate estimates to help real estate managers know about the prophecies. Buying a home will be a long-term goal for most people but there are many people who make big mistakes when buying real estate. One of the common mistakes is buying expensive buildings but not worth it. Different methods have been used in the value.

The purpose of this statistical analysis is to help us understand the relationship between the features of the house and how these conditions are used to predict the price of the house. It uses Regression algorithms comparisons to find the appropriate model to predict house price. Therefore, it may be helpful for people to avoid making mistakes. The results have proven that this method produces less error and greater accuracy than the single algorithms used. The goal of this project is to create a machine learning model that can accurately measure the value of a given home.



Unsupervised Learning Method:

In unsupervised learning we don’t have to direct the model. It predominantly manages the un-labelled information. Unsupervised learning algorithms include anomaly detection, clustering, neural networks, etc.



II.LITERATURE REVIEW

A browser-based application intended for real estate. The proposed system is a standard system that can be accessed in different locations. System over comes with all the different barriers we have in the existing system and comes up with a solution. The main purpose of the system is to predict the price of a home. The proposed system uses a machine learning method to predict the price. We use a supervised reading method to predict. From downloaded online training data sets, data collected from sources such as Kaggle, Boostan UCI ML Repository and global data websites. We use many parameters to predict the price of a house. The proposed system is an application based on an Internet browser. The real-time app where we can access the app is in a different location. The main purpose of this program is to predict the price of a home. The proposed system uses parameters such as house size, balcony, bathroom number, location and other parameters to predict house price. The system uses a line regression model to predict the price. We use effective price predictors. The system helps wealth to make quick decisions.

This paper estimates the changes in the house pricing. Housing price is strongly correlated to other factors such as location, area, population etc. In this paper applies both traditional and advance machine learning techniques and will discuss the result of this different techniques. A dataset named “Housing Price in Beijing” is used. In this paper we will discuss machine learning techniques like Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Hybrid Regression and Stacked Generalization.

In this paper, machine learning technique Random Forest is used to build the machine learning model for house price prediction. As we know when it comes to predicting pricing of house a lot of factors affect the price. In this paper we study that Random Forest works better than benchmark model linear regression. By including features such as zip code, longitude and latitude, which are not linearly related to house price, we found that random forest model performs much better and captures the hidden information in those features. We will use data of North Virginia house price.

In this paper, house price prediction is based on historical data. The goal of this study is through analyzing a real historical transactional dataset to derive valuable insight into the housing market in Melbourne city. In this paper different machine learning techniques are used like Linear Regression, Polynomial Regression, Regression Tree, Neural Network, and SVM. In this paper “Melbourne Housing Market” dataset is used.

In this paper there are three factors that are considered for the price of a house which includes physical conditions, concepts and location. The objective of the paper is prediction of residential prices for the customers considering their financial plans and needs. This examination means to predict house prices in the city with Linear Regression. Linear Regression will predict the exact numerical target value unlike other models which can only classify the output. MAE, MSE, RMSE are used to check the quality of model.

Algorithm Studied:-

Linear Regression:-

Linear regression is a supervised learning technique. It is responsible for predicting the value of a dependent variable (Y) based on a given independent variable (X).

Equation:

$Y = mX + b$

Y is dependent variable.

X is independent variable.

Multiple Linear Regressions:-

Multiple Linear Regression a new version of the linear regression which is more powerful which works with the multiple variables or the multiple features it helps to predict the unknown value of the attribute from the known value of the two or more attributes which will be also known as the predictors.

LASSO Regression:-

LASSO stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is Vast enough to improve those inclination of the model on over-fit. Least ten variables can foundation over fitting and Huge enough will cause computational tests.

Decision Tree:-

Decision Tree is a tool, which can be employed for Classification and Prediction. It has a tree shape structure, where each internal node represents test on an attribute and the branches out of the node denotes the test outcomes. Once the Decision Tree is formed, new instances can be classified easily by tracing the tree from root to leaf node. Classification through Decision Tree does not require much computation. Decision Trees are capable of handling both continuous and Categorical type of attributes.

Random Forest:-

RF is a regression technique that combines the performance of numerous DT algorithms to classify or predict the value of a variable. That is when RF receives an (x) input vector, made up of the values of the different evidential features analyzed for a given training area, RF builds a number K of regression trees and averages the results.

Neural Network:-

As a neural network in our brain, ML neural network also contains neurons, synapses, and layers. A neural network contains an input layer — a set of input features. Also, it usually contains one or more hidden layers. Each layer contains some number of nodes as neurons, and links as synapses. The last layer called “output layer” is the layer with answers.

Proposed System:-

As we know there are many ML algorithms that can be used for house price prediction. Every algorithm has its advantages and disadvantages. So, we will use ensemble learning to build our system. In ensemble learning we can combine set of individual learners (base model) together and build our final model. Our main purpose of combining different base model is to improve prediction and achieve higher accuracy. Any machine learning algorithm can be base model such as linear regression, decision tree, random forest etc.

Step 1: Load the dataset.

Step 2: Data Pre-processing contains data cleaning, data editing, data reduction. Data cleaning is process where inaccurate data or if a data field is empty, then value is filled using mean or median or entire record is deleted from data. Data editing is process where outliers are picked from data and eradicated. Data reduction is termed as the process of reducing data using some kind of normalization for easy process of data.

Step 3: Determine the Dependent and Independent variables. In our dataset price will be the dependent variable. And Independent variables will be Area, Location, No. of Bedrooms etc.

Step 4: Split dataset into training set and testing set. Training dataset will be used to train the model and testing dataset will be used to test the model.

Step 5: Training and Testing dataset will be given to different base models. Any machine learning algorithm can be base model. For example, linear regression, decision tree, random forest etc.

Step 6: Voting will select the best model for house price prediction. The model which gives highest accuracy and lowest error rate will be chosen.

Step 7: Final selected model will be given the training and testing dataset to predict the price.

Step 8: To evaluate the performance of the model we will use different measure of errors like Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE).

1. Data collection and analysis:

My process can be divided into several stages. The first stage is the data collection phase where I collected the data online. This will be used to train the machine learning model. The data collected in this category is raw and unstructured data.

According to the data the price column in the database is a dependent variable and some columns are independent variables (also called features).

- This data set has many features, including:
- Estimated income in the area of the house
- Average number of total rooms in the area

2. Data Cleaning:

Missing data is always a problem in real life situations. Areas such as machine learning and data mining face serious problems in predicting the accuracy of their models due to low data quality caused by missing values. In these areas, loss treatment is a key point to focus on making their models more accurate and relevant. To clean up data, I check to see if there are any missing values in any green data lines. However, in my data set no empty lines were found. So, I moved on to the next phase which is pre-data processing.

3. Data Processing:

I have converted my raw data into an organized form to fit a machine learning model. Since I have to use a multivariate regression model and have to be trained by my database, it is necessary that all independent variables store information in the form of numbers and not text. After manually collecting data by scratching the web, there may be some errors in the input, empty or empty values, human errors or unrealistic values that we call outliers.

So, in order to overcome this inconsistency, we need to process and refine the data from these aggregate values. There is a great need for Pre-Data Processing because if the data we provide to our model is accurate and flawless, only the model will be able to provide accurate estimates that are very close to the actual value. In Pre-Data Processing and Cleaning, we remove empty values, review all databases and also remove unnecessary data columns (independent attributes) for the purpose of precision.

4. Data Visualization:

Data viewing is a data analysis field related to visual representation of data. It organizes data clearly and is an effective way of imagining communication from data. We use data view, we can get a visual summary of our data. With pictures, maps and graphs, the human mind has an easy time to process and understand any given data.

5. Distribution of Data:

Selecting the size of each data set may be somewhat speculative and dependent on the entire sample size, and the full discussion is not in the program for this series. In general, additional training data results in better model performance and potential performance, and additional test data results in greater analysis of model performance in general.

6. Training regression model:

To model the model, 80% of the database was used and to test the 20% database model was used. From the picture above, we can see that the value depends on many factors and these factors are different conditions / features. The model function will be to calculate the coefficients and to calculate the 'c' interval. After calculating these, the model will be able to calculate the price of any custom inputs. Row line creates a number where you enter your given numbers (X) and output the target variable we want to find (Y).

7. Model testing:

To test how well the model works, I created a scattering area that shows comparisons between actual real estate prices in the database and prices predicted by the model. From the figure below, it can be concluded that in some datapoints, the actual price is very close to the predicted value which means that in some datapoints the model is more accurate. However, this figure also shows that in some cases the difference between the actual price and the predicted price is large which indicates that in some data the result is not very accurate. All in all, we can say that the model has the right amount of accuracy. Here is the code: predictions = model.predict (x_test) The prediction variable contains the estimated values of the elements stored in the x_test. Since we used the train_test_split method to store the actual values in y_test, what we want to do next is to compare the same predictive member values with y_test values. Here is the integrated structure generated by this code: calculation method that means square error. Fortunately, it doesn't really need to. Since the root means square error it is just a square root of the square root error. After selecting my best model, I went on to evaluate its effectiveness by calculating the square root of error (RMSE) and comparing actual values with predicted values.

8. Test evaluate the performance of our mode

- It means a complete mistake
- Means a square error Root means square error Unlike total error and square error, scikit-learn does not actually have a built-in root

III.RESULT

The retrospective model built into this post is incomplete, able to calculate approximately 87.8% of the local Retail Price variance. To better predict the values outside this range, some changes can be made to the model. Improvement ideas include features for log conversion and coding of class variables as usual. In this paper, I used a regression line model to make a prediction.However line reversing works very well with 84.5% accuracy. Thus, the model selected for this paper (line reversal) has the highest accuracy. This is to understand the flow of the algorithm, this algorithm works well in n number of parameters. In our project we are writing a program that works for flexible data (means the number of n parameters and the number of records n) number of parameters.

IV.CONCLUSION

The demand for commercial housing is generally divided into self-occupied demand, investment demand, and speculative demand. Self-occupation demand is self-occupation investment demand is to buy commercial housing and rent it out to obtain rental income; speculative demand buys commercial housing in anticipation of rising house prices and sells it after the price rises; and the purpose is to earn the price difference. The demand for self-occupation and investment is the supporting force of the commercial housing market. In particular, the demand for self-occupation has been encouraged and supported by policies. In addition to driving up housing prices, speculative demand squeezes out part of the demand for owner-occupiers and blows up the housing market bubble, which is harmful to the commercial housing market. This article combines the generalized linear regression model to build a real estate price prediction model. Through the simulation and comparison experiments, it can be seen that the housing price forecasting system based on the generalized regression model proposed in this article has a high housing price forecasting accuracy.

In this paper, I have used the Linear retreat model to predict the price of different homes. It comes under the supervision of a supervised learning environment which is another form of machine learning. All the steps required to successfully complete the housing pricing system have been completed. It is evident that the listing of the regression mode is appropriate for the purpose of forecasting housing prices. Our goal is achievable as we have successfully implemented all our boundaries as stated in our Goals column. It is evident that the circle level is a very effective indicator of real estate prices and that Linear Regression is the most effective model of our Dataset with RMSE School.

This study helps us to discover assets and liabilities of different machine learning models. As we know machine learning has plenty of algorithms that can be used for house price prediction. The existing systems focuses on single models only. We proposed to use multiple different model which can be used for prediction and focuses on more accurate results. We proposed to use ensemble learning method as it has capability of combining multiple ml models will help us discover different aspects of data. Hence, this methodology is anticipated to give higher accuracy compared to other single models.

Future Enhancements:

Our model had a good accuracy score, but there is still room for improvement. In real world scenario, we can use such a model to predict house prices. This model should check for new data, once in a month, and incorporate them to expand the dataset and produce better result.

Reference

1. *B. Yang and B. Cao, "Research on ensemble learning-based housing price prediction model," Big Geospatial Data and Data Science, vol. 1, no. 1, pp. 1–8, 2018.*
2. *J. Q. Guo, S. H. Chiang, M. Liu, C. C. Yang, and K. Y. Guo, "Can machine learning algorithms associated with text mining from internet data improve housing price prediction performance?" International Journal of Strategic Property Management, vol. 24, no. 5, pp. 300–312, 2020.*
3. *J. M. Montero, R. Minguez, and G. Fernandez-Aviles, ´ "Housing price prediction: parametric versus semi-parametric spatial hedonic models," Journal of Geographical Systems, vol. 20, no. 1, pp. 27–55, 2018.*
4. *A.R. A. Yakub, M. Hishamuddin, K. Ali, R. B. A. J. Achu, and A. F. Folake, "The effect of adopting micro and macro-economic variables on real estate price prediction models using ann: a systematic literature," Journal of Critical Reviews, vol. 7, no. 11, pp. 492–498, 2020.*
5. *L. Yu, C. Jiao, H. Xin, Y. Wang, and K. Wang, "Prediction on housing price based on deep learning," International Journal of Computer and Information Engineering, vol. 12, no. 2, pp. 90–99, 2018.*
6. *J. Lee and J. P. Ryu, "Prediction of housing price index using artificial neural network," Journal of the Korea Academia-Industrial cooperation Society, vol. 22, no. 4, pp. 228–234, 2021.*
7. *R. Liu and L. Liu, "Predicting housing price in China based on long short-term memory incorporating modified genetic algorithm," Soft Computing, vol. 23, no. 22, pp. 11829–11838, 2019.*
8. *S. Muralidharan, K. Phiri, S. K. Sinha, and B. Kim, "Analysis and prediction of real estate prices: a case of the Boston housing market," Issues in Information Systems, vol. 19, no. 2, pp. 109–118, 2018.*
9. *K. S. Yoon, J. M. Lee, S. J. Ko, H. J. Kim, and J. H. Kim, "Analysing impact of price ceiling system on housing market using machine learning," Journal of the Architectural Institute of Korea, vol. 37, no. 8, pp. 221–228, 2021.*
10. *Y. R. Lin and C. C. Chen, "House price prediction in taipei by machine learning models," International Journal of Design, Analysis and Tools for Integrated Circuits and Systems, vol. 8, no. 1, pp. 89–94, 2019.*
11. *M. Ozdemir, K. Yildiz, and B. Buyuktanir, "Housing price estimation with deep learning: a case study of sakarya Turkey," Bilecik Seyh Edebali Universitesi Fen Bilimleri Dergisi, vol. 9, no. 1, pp. 138–151.*
12. *J. Hong, H. Choi, and W. S. Kim, "A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea," International Journal of Strategic Property Management, vol. 24, no. 3, pp. 140–152, 2020.*
13. *C. Li, H. Zhu, X. Ye et al., "Study on average housing prices in the inland capital cities of China by night-time light remote sensing and official statistics data," Scientific Reports, vol. 10, no. 1, pp. 7732–7750, 2020.*
14. *J. H. Chen, T. Ji, M. C. Su, H. H. Wei, V. T. Azzizi, and S. C. Hsu, "Swarm-inspired data-driven approach for housing market segmentation: a case study of Taipei city," Journal of Housing and the Built Environment, vol. 36, no. 4, pp. 1787–1811, 2021.*
15. *D. Cao and X. Tian, "Raw anode volume density prediction algorithm based on the genetic algorithm," SN Computer Science, vol. 3, no. 5, pp. 354–372, 2022.*