International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Prediction Of Data Analysis Using Machine Learning Techniques

Dr.M.Balakrishnan[1], Dr.A.B.Arockia Christopher[2], Dr.A.S.Muthanantha Murugavel[3], J.Ramprasath[4]

[1,2]Assistant Professor (SG), Department of Information Technology, Dr. Mahalingam College of Engineering and Technology, Pollachi, Coimbatore,Tamilnadu, India.
[3]Associate Professor, Department of Information Technology,Dr. Mahalingam College of Engineering and Technology, Pollachi, Coimbatore,Tamilnadu, India.
[4]Assistant Professor,Department of Information Technology,Dr. Mahalingam College of Engineering and Technology, Pollachi, Coimbatore,Tamilnadu, India.

Abstract: The rising demand for prediction has made it more popular and a helpful tool. Thus most of us go for prediction most of the time. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Predicting housing prices with real factors is the main crux of our research project. Here we aim to make our evaluations based on every basic parameter that is considered while determining the price. We use various machine learning techniques in this pathway, and our results are not sole determination of one technique rather it is the weighted mean of various techniques to give most accurate results. The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied. The environment and surroundings are also important features to take in mind to predict its value. But most of us will only take the building and its physical materials for consideration but the real up and down of the price values depend on those extra surrounding environments which are placed nearby. The main role of the idea is not only to predict the price but also to find some of the possibilities after the prediction has been done successfully.

Keywords – Property prediction, Find the possibilities of new insights, Machine Learning methods.

1. INTRODUCTION

We are predicting the sale price of the houses using various machine learning algorithms. Housing sales price are determined by numerous factors such as area of the property, location of the house, material used for construction, age of the property, number of bedrooms and garages and so on. This paper uses machine learning algorithms to build the prediction model for houses. Here, machine learning algorithms such as logistic regression and support vector regression, Lasso Regression technique and Decision Tree are employed to build a predictive model. Real estate is a thriving and appealing investment sector not only in our country but throughout the world. It is one of the wealth measures for the country and also for a person.

It is an important problem for all the stakeholders like house buyers, house owners, agents, real estate brokers, investors. The price of the property will keep on increasing unevenly. Thus, the price value will depend on certain features and criteria like the area of the property [sq. ft], the number of rooms and balconies, and even more. The location where the property has been situated plays a vital role in prediction.

Some additional features are also able to decide the price value of the property like the availability of the basic needs and people review. These features will change over time and it is not a constant one. The above features are changing properties which predict a variable output through time and surroundings.

Thus the prediction will become more accurate when we have all these data in a correct format and most probably true values. The main role in prediction is data collection and data cleaning. It will help us to have the true data values and also be able to change the unnecessary data into either useful data or true data. So the prediction of the price will depend on multiple features. These all features contribute their property to predict the single output called price. The Property prediction factors will depend upon some of the nearby features which are shown in figure 1.

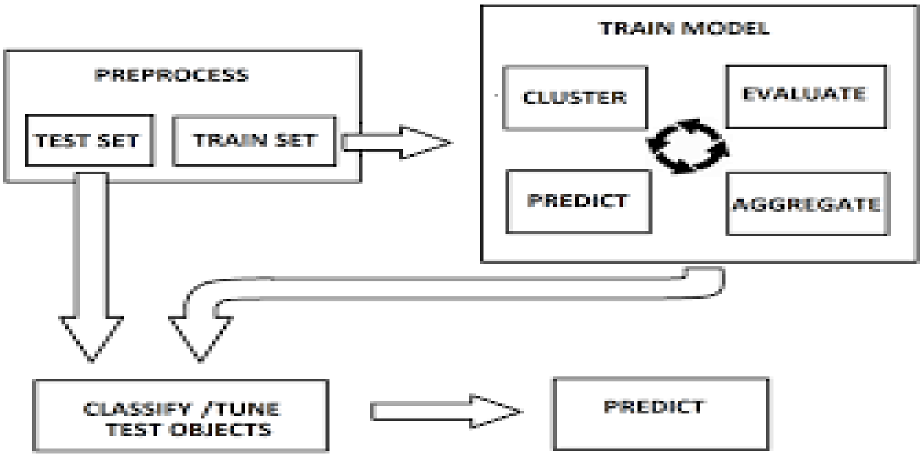


Figure 1: Predicting house prices using machine learning

2. LITERATURE WORK

Many works are related to our implementation. But the main part of our implementation is to make use of the predicted results in a much efficient way. The results are dependent on some independent features. The predicted output can also be varied by some other related data. Thus in our implementation, the price of a house or residential area can be predicted by using machine learning concepts whereas those predicted values are not accurate. They can also be able to change over time by some additional features like the surroundings and current issues. Many algorithms, such as Linear Regression, Random Forest, and others are used to forecast home prices.

This paper was proposed by a group of students who believe that combining linear and boosted algorithms with neural networks would improve prediction accuracy. Here various features of data, which are very essential for prediction, are used. Initially, it was not cleaned but after they have been cleaned the data well for better performance. They have used some of the regression algorithms in this implementation. The dataset was run through the algorithms of the neural network and will take all the results of those algorithms as feedback to it. It will analyze the information and present the results. Thus the Boosted Regression with a neural network is to improve accurateness.

Real Estate Property is not only the basic need of a man but today it also represents the riches and prestige of a person. Investment in real estate generally seems to be profitable because their property values do not decline rapidly. Changes in the real estate price can affect various household investors, bankers, policy makers and many. Investment in real estate sector seems to be an attractive choice for the investments. Thus, predicting the real estate value is an important economic index. India ranks second in the world in number of households according to 2011 census with a number of 24.67 crore. India is also the fastest growing major economy ahead of China with former's growth rate as 7% this year and predicted to be 7.2% in the next year.

The real estate market is exposed to many fluctuations in prices because of existing correlations with many variables, some of which cannot be controlled or might even be unknown. Housing prices can increase rapidly (or in some cases, also drop very fast), yet the numerous listings available online where houses are sold or rented are not likely to be updated that often. In some cases, individuals interested in selling a house (or apartment) might include it in some online listing, and forget about updating the price. In other cases, some individuals might be interested in deliberately setting a price below the market price in order to sell the home faster, for various reasons. In this paper, we aim at developing a machine learning application that identifies opportunities in the real estate market in real time, i.e., houses that are listed with a price substantially below the market price. This program can be useful for investors interested in the housing market. We have focused in a use case considering real estate assets located in the Salamanca district in Madrid (Spain) and listed in the most relevant Spanish online site for home sales and rentals. The application is formally implemented as a regression problem that tries to estimate the market price of a house given features retrieved from public online listings

ALGORITHMS USED

Linear Regression Algorithm:

It is one of the types in supervised Learning, Where we will be having both the X and Y form in which we can the Y->output for the given X as input from the dataset. Linear Regression uses the mathematical formula Y=mx +c, this equation gives the best fit line going through the data points. However the data points are scattered so, it's not possible to draw the perfect line but draws a kind of best fit line. Simple linear regression and multiple linear regression are the two forms of linear regression. Simple Linear is used to forecast a value using only one independent variable, whereas Multiple Linear is used to predict a value using one or more independent variables. The Life Cycle of the Machine Learning is shown in the below figure 2.

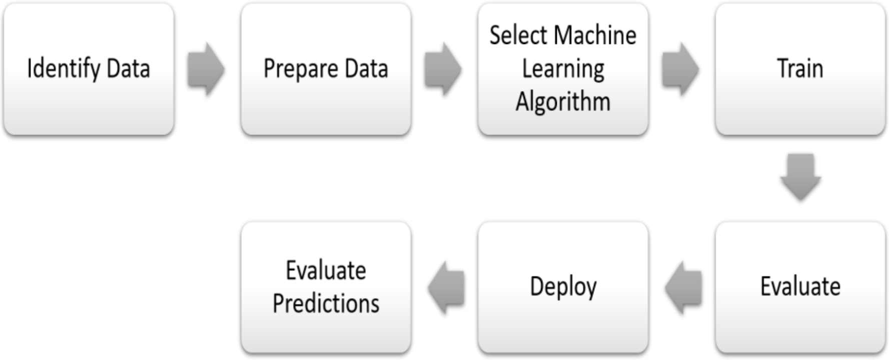


Figure 2: Building a machine learning model

Correlation Matrix:

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. To summarize a large amount of data where the goal is to see patterns. In our example above, the observable pattern is that all the variables highly correlate with each other.

To input into other analyses. For example, people commonly use correlation matrixes as inputs for exploratory factor analysis, confirmatory factor analysis, structural equation models, and linear regression when excluding missing values pairwise. As a diagnostic when checking other analyses. For example, with linear regression, a high amount of correlations suggests that the linear regression estimates will be unreliable.

Mean Absolute Error (MAE):

Given any test data-set, Mean Absolute Error of your model refers to the mean of the absolute values of each prediction error on all instances of the test data-set. Prediction error is the difference between the actual value and the predicted value for that instance. Statistically, Mean Absolute Error (MAE) refers to the results of measuring the difference between two continuous variables. Let's assume variables M and N represent the same phenomenon but have recorded different observations.

For a given scatter plot of x points, where point j has coordinates (Mj, Nj). Our Mean Absolute Error (MAE) will be the average vertical distance between each point and the N=M line. This is also known as the One-to-One line. MAE will also at this point be the average of total horizontal distance between each point and the N=M line.

Root mean squared error (RMSE):

Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance. To compute RMSE, calculate the residual (difference between prediction and truth) for each data point, compute the norm of residual for each data point, compute the mean of residuals and take the square root of that mean. RMSE is commonly used in supervised learning applications, as RMSE uses and needs true measurements at each predicted data point.

Root mean square error can be expressed as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}||y(i) - \hat{y}(i)||^2}{N}},$$

Where N is the number of data points, y(i) is the i-th measurement, and y ̃(i) is its corresponding prediction.

Mean squared error (MSE):

One of the most common metrics used to measure the prediction accuracy of a model is MSE, which stands for mean squared error. It is calculated as:

MSE = (1/n) * Σ(actual – prediction) 2

Where:

● Σ –a fancy symbol that means"sum"

n – sample size

actual – the actual data value
prediction – the predicted data value The lower the value for MSE, the more accurately a model is able to predict values.
The lower the value for MSE, the more accurately a model is able to predict values.
Variance Score:
In linear regression, variance shows the difference of scores between the actual and the predicted values, this can be also taken as the accuracy of the model.

Mean Relative Error (MRE):
To compute the MRE we start by computing the Relative Error (RE) for each data set. The RE of a data set is computed in this way

$$\mathrm{RE}_i = \sum_{j=1}^{n_i} \frac{w_{ij}}{w_i} \frac{|p_{ij} - d_{ij}|}{|d_i|}$$

for

$$w_i = \sum_{j=1}^{n_i} w_{ij} > 0$$

where i refers to the data set and j to a given point in data set i. dij stands for the data, pij for the model prediction and wij for the associated weight coefficient. Finally, di in the denominator represents the average of all data points in set i (dij):

$$d_i = \frac{1}{n_i} \sum_{j=1}^{n_i} d_{ij}$$

In the case the sum of the weights wi is equal to zero we simply take the relative error to be zero. The implication is that this data set does not contribute to the MRE.

MRE is then the mean of all REi with wi greater than 0.

$$\mathrm{MRE} = \frac{1}{n'} \sum_{i=1}^{n} \mathrm{RE}_i$$

where n' is the number of data sets with wi greater than 0.

DATASET

Data Description

In our research, we have used a large amount of data which are more essential for the contribution of the price prediction. The features which have been used in our projects are the area of the property, no. of bathrooms, no. of bedrooms, balcony, and area type. We have used more than 10000 rows of data which consist of around 8 feature columns that are going to take part in the price prediction. The initial data we collected may or may not consist of false data and null values. Hence, the first and foremost process is to clean the data for better understanding.

When the data is cleaned well it will help us to predict the value with high accuracy. Data cleaning can be one of the time-consuming processes but it will lead the remaining process safely and efficiently. Thus every feature has its priority and contribution in prediction. The high priority feature has more impact on the output whereas the less priority feature will not have that much impact on the output of the prediction than the high priority feature. After cleaning of data, the prediction process will take place. The importance of the features is based on the proposal we implemented which was shown in figure 3. The extra data which is gathered for finding further insights and possibilities also wants to be cleaned. But these data are time-series data, they keep on changing to time. Hence these data are managed frequently to make the process more elegant.

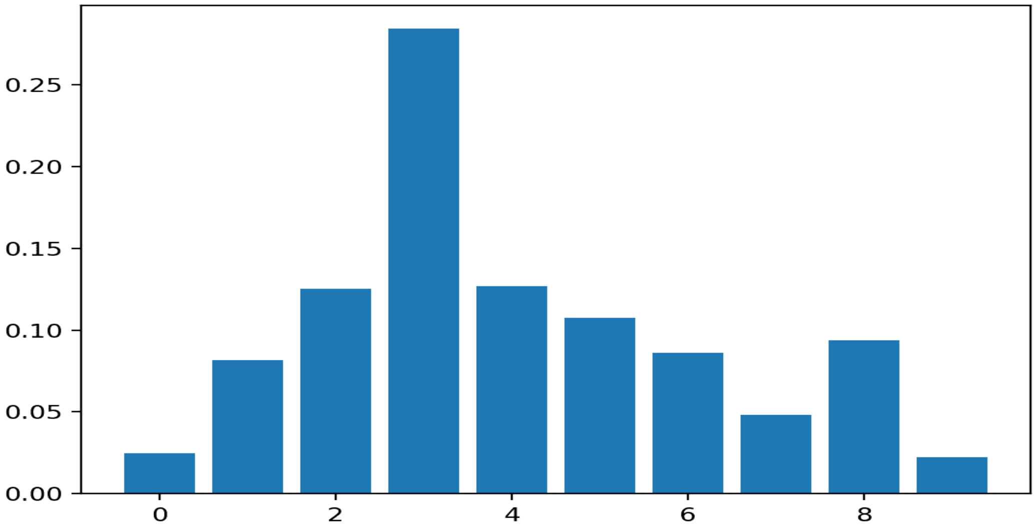
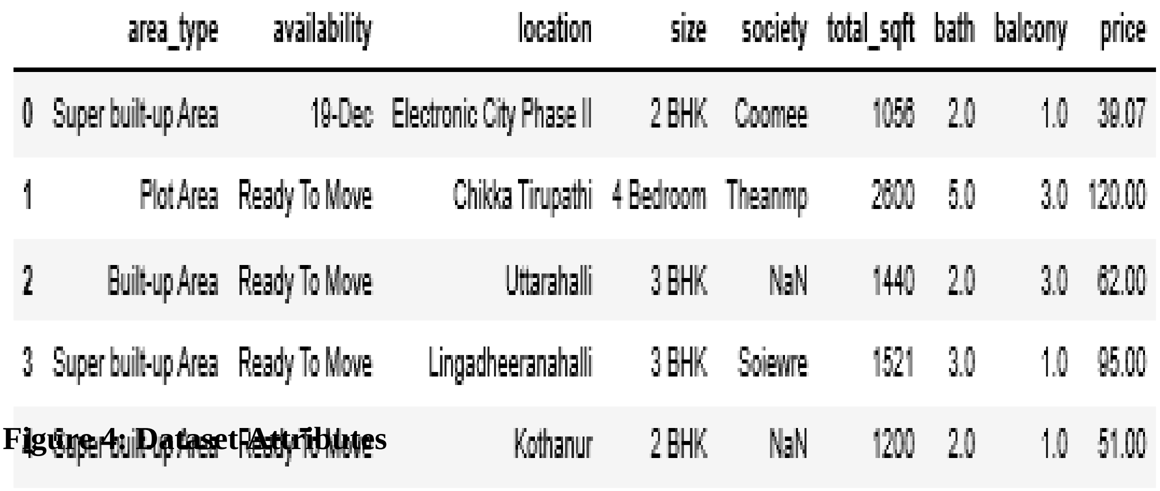


Figure 3: Calculate Feature importance



	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

Figure 4: Dataset Attributes

Prediction Analysis:

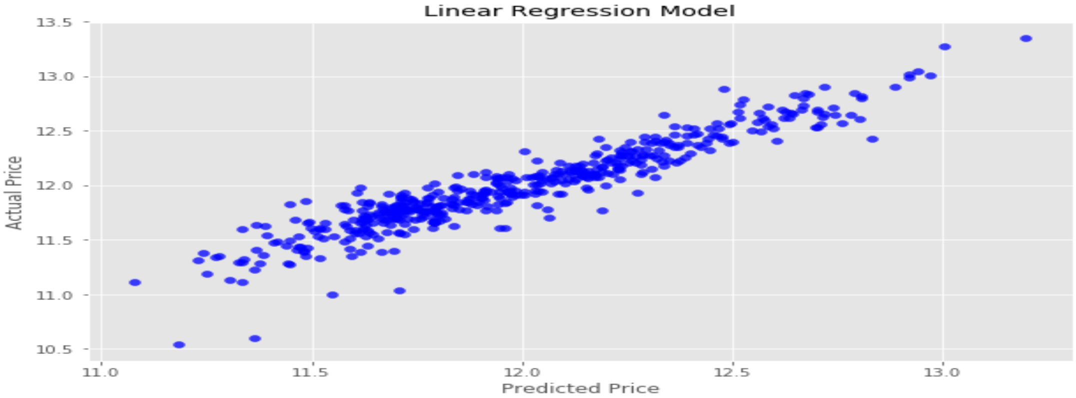


Figure 5: Actual Price Vs Predicted Price

3. CONCLUSION

By summarizing, Property and residential price is dependent upon the features which are all needed for the property. The final prediction of the model will be used to make new insights for upcoming scenarios. The idea of creating a useful model by using the efficient algorithm will make the implementation fall on the higher-end whereas the final predictions are used for further future uses and possibility findings will also make it extend to a higher user level. Hence, our main perspective is to make use of prediction to make a new model and make it the beginning of a new idea.

Thus the implementation is not based on the accuracy of the prediction, whereas it is based on how we are going to use those predictions in a better way. Hence, In the future, there are a lot of opportunities for new challenges. The first main challenge is about data gathering for additional insights. Thus the data gathering is more time-consuming work and also a slow process. There are some of the open data for quick use but it may or may not be useful for the model which we are going to showcase. Thus data gathering is one of the most important challenges in the future. The data are kept on increasing exponentially. The data which we have collected will expire at some point. The retention of the data keeps on decreasing in the upcoming days. The other challenges are based upon the type of work we are going to use the model. The new idea of using the predicted value differs for every individual. The efficient idea will lead to an efficient and better working model in the future.

4. REFERENCES

[1] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth EmaFebrita, Wayan Firdaus Mahmudy, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization",(IJACSA)International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017.

[2] Al-Alawi, S.M., AbdulWahab, S.A., &Bakheit, C.S. (2008)."Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone." Environmental Modelling and Software, 23(4), 396-403. DOI:10.1016/j.envsoft.2006.08.007.

[3] Alejandro Baldominosetal. "Identifying Real Estate Opportunities Using Machine Learning". In: MDPI Applied Sciences (Nov.2018).

[4] Analysis And Prediction Of Real Estate Prices: A Case Of The Boston Housing Market Sharmila Muralidharan, Seattle University, muralidh@seattleu.edu Katrina Phiri, Seattle University, phirik@seattleu.edu Sonal K. Sinha, Seattle University, sinhas1@seattleu.edu Ben Kim, Seattle University, bkim@seattleu.edu

[5] Ayush Varma, Abhijit Sharma, Sagar Doshi, Rohini Nair, "House Price Prediction Using Machine Learning and Neural Networks", INSPEC number 18116205, April 2018.

[6] Gu Jirong, Zhu Mingcang, and Jiang Liuguangyan. (2011).Housing price based on genetic algorithm and support vector machine". In: Expert Systems with Applications 38 pp. 3383–3386.

[7] James Frew and G. Donald Jud: "Estimating the Value of Apartment Buildings" uses the Hedonic model for predicting the price with higher variation in price.

[8] Lu. Sifei et al, A hybrid regression technique for house prices prediction. In Proceedings of IEEE Conference on Industrial Engineering and Engineering Management: 2017. Machine Learning and Neural Networks", INSPEC number 18116205, April 2018.

[9] Neelam Shinde, Kiran Gawande, "Valuation of House Price Using Predictive Techniques", International Journal of Advances in Electronics and Computer Science,ISSN: 2393-2835(IJAECS), Volume - 5, Issue - 6, June -2018.

[10] Park, Byeonghwa, and Jae Kwon Bae. "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data." Expert Systems with Applications 42.6 (2015):2928-2934.

[11] Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques, Expert Systems with Applications 42(1): 259–268.

[12] Rochard J. Cebula (2009).The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District; The Review of Regional Studies 39.1 (2009), pp. 9– 22.

[13] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh, "A Hybrid Regression Technique for House Price Prediction", December 2017.

[14] Visit Limsombunchai, Christopher Gan and Minsoo Lee, "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network", American Journal of Applied Sciences 1 (3):193-201, 2004, ISSN 1546-9239, 193 – 201.

[15] Yang Yonghui. Research on HousePrice Prediction Based on MultiDimensional Data Fusion [J]. International Journal of Advanced Network, Monitoring and Controls Volume 05, No.01, 2020. Xi'an Technological University Xi'an, 710021, China.