

# House Price Prediction Using Machine Learning

Prof. Archana lomte<sup>1</sup>, Nikita Bhagwat<sup>2</sup>, Kaveri Adhav<sup>3</sup>, Trupti bhavar<sup>4</sup>

<sup>1, 2, 3, 4</sup> Dept of Computer Engineering

<sup>1, 2, 3, 4</sup> BSIOTR, Pune, Maharashtra, India

**Abstract-** The development of the falling or rising of the house prices has attracted interest from the person in addition as many different interested parties. There square measure many previous researches works that used varied regression techniques to trot out the question of the changes house value. This work considers drawback of fixing house value as a classification drawback and applies machine learning techniques to predict whether or not or not house prices will rise or fall. This work applies varied feature selection techniques like variance influence issue, info price, principal part analysis and knowledge transformation techniques like outlier and missing value treatment additional as box-cox transformation techniques. The performance of the machine learning techniques is measured by the four parameters of accuracy, precision, specificity and sensitivity. The work considers a pair of distinct value zero and one as varied classes. If the value of the class is zero then we tend to tend to admit that the value of the house has pale and if the value of the class is one then we tend to tend to admit that the value of the house has raised.

**Keywords-** Machine Learning, Linear Regression, House price prediction.

## I. INTRODUCTION

Development of civilization is that the muse of increase of demand of homes day by day. correct prediction of house prices has been unendingly a fascination for the patrons, sellers and for the bankers in addition many researchers have already worked to unravel the mysteries of the prediction of the house value. There area unit a unit many theories that area unit born as a consequence of the analysis work contributed by the numerous researchers everyplace the earth. variety of those theories believe that the geographical location and culture of calicular} house verify but the house prices will increase or decrease whereas their area unit a unit totally different school of thought world organization agency emphasize the socio-economic conditions that for the foremost part play behind these house value rises. we have a tendency to tend to any or all acknowledge that house value is also a spread from some made public assortment, so clearly prediction of prices of homes is also a regression task. To forecast house value one person generally tries to seek out similar properties at his or

her neighborhood and supported collected info that person will decide to predict the house value.

Machine learning (ML) is a district of AI that works manually victimization trained model input Paid throughout the coaching amount. Machine learning is employed to perform tons of computing tasks. it's additionally accustomed create predictions with the utilization of computers. Machine learning is typically additionally accustomed devise advanced models. The principal value of machine learning is to allow the PCs to be told things naturally while not the help of individuals. Machine learning is incredibly helpful and is wide used round the whole world. the method of machine learning involves providing information and so coaching the computers by building machine learning models with the assistance of assorted algorithms. A set of machine learning is closely associated with procedure statistics, that focuses on creating predictions victimization computers; however not all machine learning is applied mathematics learning. The study of mathematical optimization delivers ways, theory and application domains to the sphere of machine learning. data processing may be a connected field of study, that specialize in explorative information analysis through unsupervised learning. In its application across business issues, machine learning is additionally mentioned as prognostic analytics.

## II. EXISTING SYSTEM

There square measure many approaches which will be accustomed verify the worth of the house, one amongst them is that the prediction analysis. the primary approach could be a quantitative prediction. A quantitative approach is associate degree approach that utilizes timeseries information [5]. The time-series approach is to seem for the link between current costs and prevailing costs. The second approach is to use rectilinear regression supported hedonistic rating. Previous analysis conducted by Gharehchopogh victimization rectilinear regression approach get zero,929 errors with the particular value. In rectilinear regression, determinant coefficients typically victimization the smallest amount sq. technique, however it takes an extended time to urge the most effective formula. Particle swarm improvement (PSO) is planned to search out the coefficients geared toward getting best result. Some previous researches like Marini and Walzack show that PSO gets higher results than alternative hybrid

ways. There square measure many benefits of PSO, within the little search house PSO will do higher answer search. though the PSO international search is a smaller amount than best, however on the improvement downside the price of the variable on the regression of y on x will realize a most answer victimization PSO.

### III. PROPOSED SYSTEM

Our dataset contains of assorted essential parameters and data processing has been at the foundation of our system. we tend to ab initio clean up our entire dataset and conjointly truncated the outlier price. Further, we tend to weighed every parameter supported its importance in determinant the evaluation of the system and this semiconductor diode US to extend the value that every parameter withholds within the system. we tend to shortlisted three totally {different|completely different} machine learning algorithms and tested our system with different mixtures that may guarantee best presumably reliableness of our results.

Even afterward, we tend to followed a novel approach to extend the accuracy, our survey semiconductor diode to a conclusion that the particular assets price conjointly depends on near native amenities like depot, grocery, school, hospital, temple, parks etc. And currently we tend to propose our distinctive approach that may counter this would like. we tend to use Google maps API and supported neighborhood search we tend to slim down on a radius of zero.5 km. currently if we discover any such public places within the circle, we tend to increase the value of the property correspondingly. we tend to carried this out with manual examples and this gave US tremendous leads to terms of accuracy in prediction.

- Algorithms used:

#### 1.LINEAR REFRESSION-

Linear regression is that the simplest methodology for prediction. It uses 2 things as variables that ar the variable quantity and also the variable that is that the most important one initial whether or not the variable quantity. These regression estimates ar want to justify the link between one variable and one or additional freelance variables. The equation of the equation with one dependent and one variable quantity is outlined by the formula.  $b = y + x*a$  wherever,  $b =$  calculable variable score,  $y =$  constant,  $x =$  parametric statistic, and  $a =$  score on the variable quantity depends upon the value of the variable quantity or variables. the best kinds of regression assume that the link between the freelance and dependent variables follows a line. The origin of the term

"regression" to explain the method of fitting lines to information is stock-still in an exceedingly study of genetic science by Sir Galton within the late nineteenth century. He discovered that fathers UN agency were very short or very tall cared-for have sons whose heights were nearer to the common height. He referred to as this development "regression to the mean".

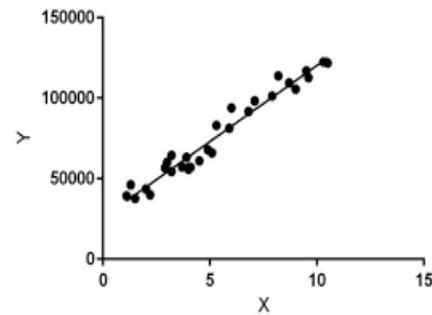


Fig 1. Linear Regression

#### UNDERSTANDING REGRESSION.

You might recall from basic pure mathematics that lines are often outlined during a slope-intercept type like  $y = a + bx$ . during this for the letter y indicates the variable and x indicates the variable. The slope term b specifies what quantity the road rises for every increase in x. Positive price outline lines that slope upward whereas negative price outline lines that slope downward. The term a is understood because the intercept as a result of it specifies the value wherever the road crosses, or intercepts, the vertical y axis. It indicates the value of y once  $x =$  zero. Regression equations model information employing a similar slope-intercept format. The machine's job is to spot price of a and b in order that the required line is best ready to relate the provided x price to the value of y. There might not perpetually be one operate that completely relates the value;therefore, the machine should even have a way to quantify the margin of error. We'll discuss this comprehensive shortly.

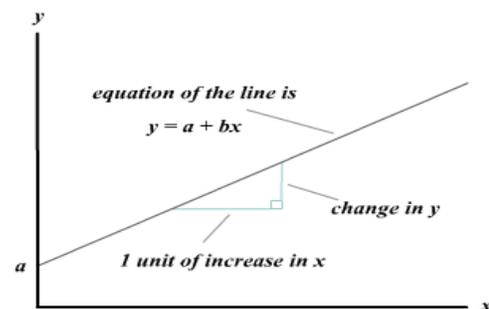


Fig 2. Regression

**IV. WORKING OF SYSTEM**

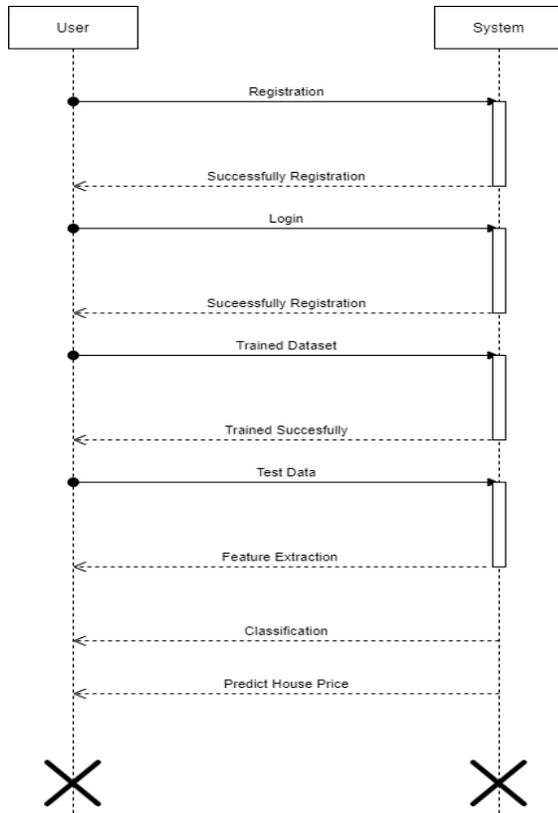


Fig 3. Sequence Diagram

The Sequence diagram higher than explains the operating of the system. The projected system is meant to be a web site with three objects namely: client, the net Interface and therefore the information Server. The information server conjointly includes the machine mechanism delineated within the rule. once the client 1st enters into the web site, they're displayed with a user interface wherever they'll enter inputs like the sort of house, the realm during which it's placed etc. a knowledge index looking then provides with outputs consisting of matching properties. Now, if the client needs to see the house value in future, they'll enter the initiate the longer term. The system can establish the date and reason it within the quarters. The rule then can cipher the value of rate and supply the results back to the client.

**V. PROPOSED TECHNIQUE**

**1. MAHINE LEARNING:**

we apply the construct of machine learning. the most aim of our analysis is to use the ideas of machine learning algorithms for making an automatic system for helmet detection. Machine Learning is that the set of AI that deals with the extraction of patterns from datasets. Machine learning is that the act of utilizing algorithms to take apart info, gain

from it, and later build a prediction or expectation concerning one thing within the world. In our new system, we tend to apply the techniques of deep learning that may be a set of machine learning rule that utilizes complicated artificial neural networks. one amongst the most areas of analysis for our projected innovation includes the usage of pre-trained real time object detection models.

**CHARACTERISTICS OF MACHINE LEARNING:**

1.The ability to perform automatic info visualization:  
A massive amount of information is being generated by businesses and customary of us on a day to day by visualizing notable relationships in knowledge, businesses cannot only build higher decisions but build confidence additionally.

2.Automation at its best:

One of the biggest characteristics of machine learning is its ability to vary repetitive tasks then, increasing productivity. an enormous vary of organizations square measure already victimization machine learning-powered work and email automation.

3.Customer engagement like never before:

For any business, one all told the foremost crucial ways in which to drive engagement, promote complete loyalty and establish sturdy consumer relationships is by triggering purposeful conversations with its target consumer base. Machine learning plays a very important role in enabling businesses and kinds to spark tons of valuable conversations in terms of consumer engagement.

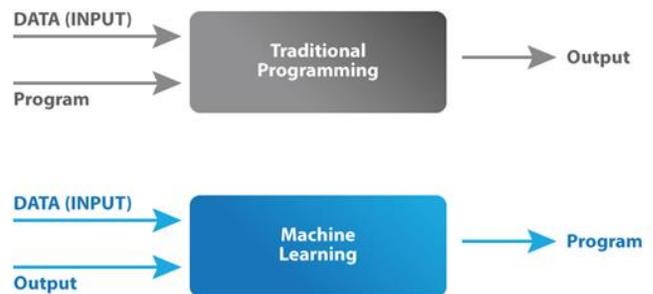


Fig 4. Machine Learning

**2.SVM:**

In machine learning, support-vector machines (SVMs, else support-vector networks) ar supervised learning models with connected learning algorithms that analyses data used for classification and statistical procedure. Developed at

AT&T Bell Laboratories by Vapnik with, it presents one in all the foremost durable prediction ways that within which, supported the mathematics learning framework or VC theory projected by Vapnik and Chervonenkis and Vapnik. Given a bunch of employment examples, every marked as happiness to a minimum of 1 or the choice of 2 classes, associate SVM employment formula builds a model that assigns new examples to a minimum of 1 class or the choice, creating it a non- probabilistic binary linear classifier. associate SVM model may be associate illustration of the examples as worths in house, mapped so the instances of the separate categories ar split by a transparent gap that's as wide as probable. New examples r then scheduled into that very same house and expected to belong to a class supported the side of the gap on it they fall. Support Vector Machine” (SVM) can be a supervised machine learning rule which can be used for every classification or regression challenges. However, it's mostly used in classification problems. among the SVM rule, we tend to tend to plot each data item as a value in n-dimensional space (where n is vary of choices you have) with the value of each feature being the value of a particular coordinate. Then, we tend to tend to perform classification by finding the hyper-plane that differentiate the two classes o.k. (look at the below snapshot).

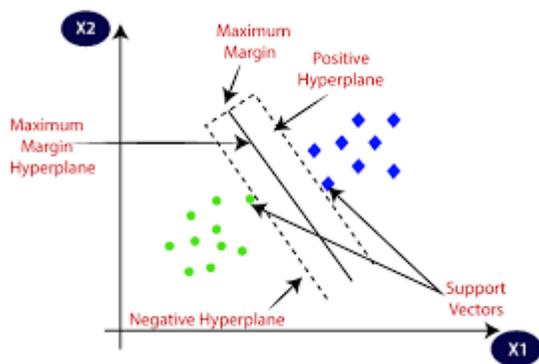


Fig 5.SVM

APPLICATION OF SVM –

- SVMs unit helpful in text and electronic text categorization, as their application can significantly cut back the need for tagged coaching job instances in every the standard inductive and transudative settings. Some ways for shallow linguistics parsing unit supported support vector machines
- Classification of images can also be performed victimization SVMs. Experimental results show that SVMs attain significantly higher search accuracy than ancient question refinement schemes once merely three to four rounds of connexion feedback. this will be to boot true for image segmentation systems, alongside those

using a modified version SVM that uses the privileged approach as suggested by Vatnik.

- The SVM formula has been wide applied among the biological and various sciences. they have been accustomed classify proteins with up to ninetieth of the compounds classified properly. Permutation tests supported SVM weights ar suggested as a mechanism for interpretation of SVM models. Support-vector machine weights have to boot been accustomed interpret SVM models among the past.

VI. SYSTEM REQUIREMENTS

1. Software Requirements:

- Windows 10
- Anaconda
- Spyder
- DB Browser (SQLite)

2.Hardware requirements:

- Processor – i5
- Hard Disk – 8 GB
- Memory – 40GB RAM

VII. UNDERSTANDING THE MATHEMATICS BEHIND SUPPORT VECTOR MACHINES

1.Length of a vector:

The length of a vector x is termed its norm, that is written as ||x||. The geometer norm formula to calculate the norm of a vector x = (x1,x2,...,xn) is:

$$||x|| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$

2.Direction of a vector:

The direction of a vector x = (x1,x2) is written as w, and is defined as:

$$w = \frac{1}{||x||} (x_1, x_2)$$

If we tend to look at figure one, we are going to see that  $\cos(\theta) = \frac{x_1}{||x||}$  and

$\cos(\alpha) = \frac{x_2}{||x||}$ . Thus, the direction vector we are going to even be

written as:

$$w = (\cos(\theta), \cos(\alpha))$$

3.Linear separability:

Linear disconnection is one necessary plan in SVM. tho' in smart cases the data will not be linearly divisible, we'll begin from the linearly divisible cases (since they are simple to grasp and deal with) and then derive the non-linearly

divisible cases. Figure half-dozen shows the two-dimensional info ar separated by a line. throughout this case, square measure saying} the data is linearly divisible. Figure seven is associate example of non-linearly divisible info, which suggests we've got a bent to cannot notice a line to separate the two-dimensional info. Similarly, for three-dimensional info, {we say we tend to ar saying} the data is linearly divisible if we are going to notice a plane to separate them.

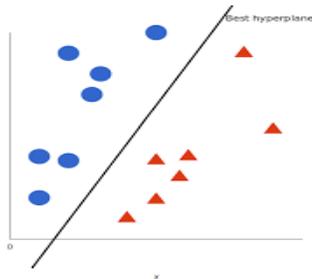


Fig 6. Two-dimensional data are separated by a line

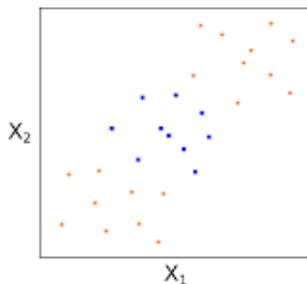


Fig 7. Non-linearly separable data

## VIII. CONCLUSION

The sales worth for the homes are calculated victimization completely different algorithms. The sales costs are calculated with better accuracy and exactness. this may be of nice facilitate for the folks. to attain these results, numerous information mining techniques are used in python language. The various factors that have an effect on the house valuation ought to be considered and work upon them. Machine learning has assisted to finish out task. Firstly, the information assortment is performed. Then information improvement is administered to get rid of all the errors from the information and build it clean. Then the information pre-processing is finished. Then with facilitate of knowledge visualized particle, completely different plots are created. This has portrayed the distribution of knowledge in several forms. Further, the preparation and testing of the model are performed. It has been found that a number of the classification algorithms were applied on our dataset whereas some weren't. So, those algorithms that weren't being applied on our house pricing dataset are born and tried to enhance the accuracy and exactness of these algorithms that were being applied on our house valuation dataset.

## REFERENCES

- [1] Lisi G. indulgent valuation models and residual house value volatility. Letters in spatial and Resource Sciences, 2019: 1-10.
- [2] Caterina Contini, Fabio Boncinelli, Caterina Romano, Gabriele Scozzafava, da Vinci Casini. value vector issue in a very selection experiment: A method proposal. Food Quality and Preference,2019,75.
- [3] Ang L I , Yi-Xiang L I , Xue-Hui L I . TensorFlow and Keras-based Convolutional Neural Network in CAT Image Recognition. 2017.
- [4] Wisanlaya Pornprakun, et al. determinant best policies for sugarcane harvest in Siam victimization bi-objective and quasi-Newton improvement ways. Advances in distinction Equations,2019, Vol.2019 (1), pp.1-15
- [5] Dayhoff J E, Deleo J M. Artificial neural networks. Cancer, 2001, 91(8):1615-1634.
- [6] Lu Lili, Hu Bin, Li Hui, Duanmu Yiting.Simulation Analysis of China's Housing value Composition and Forecast.Computer Simulation,2014,31(03):230-238.