INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 4)
Available online at: www.ijariit.com

House price prediction using various machine learning algorithms

Parth Ambalkar
parth.ambalkar@spit.ac.in
Bharatiya Vidya Bhavan's Sardar Patel Institute of Technology, Mumbai, Maharashtra

Akash Mane
akash.mane@spit.ac.in
Bharatiya Vidya Bhavan's Sardar Patel Institute of Technology, Mumbai, Maharashtra

Tanmay Maity
tanmaymaity06@gmail.com
Terna Engineering College, Navi Mumbai, Maharashtra

ABSTRACT

House price increases gradually, that is why there is a need to create such a system for prediction of house prices. This prediction will help developers knowing the selling price of a house. It will also assist customers to know about which is the perfect time to buy a flat. In this paper, we are going to predict the selling price of various houses. Selling prices are determined by various factors such as the location of the house, area of the property, the inflation rate of the current year, Apartment type, month and year of which we want to know the particular price. We are implementing various machine learning algorithms for building a predictive model for houses. We have considered housing data of 2000 properties. In this paper, we will be comparing the algorithms on the basis of parameters such as MAE, RMSE, MSE, accuracy.

Keywords— *Prediction, House price*

1. INTRODUCTION

Real Estate housing property is of great significance for the economy life. It performs not just primary demand of a human being but now it also performs richness as well as the prominence of a customer. Financing a real estate commonly seemed to a beneficial as a result their equity value does not deteriorate briskly. Diversity is real estate housing prices can disturb different ordinary investor, policy-making, banker, etc. Investment at real estate zone implies hereafter an alluring decision for investments. So, predicting a real estate amount is a crucial economic ratio. According to the 2011 census, India ranks second in a total of household work with a figure of 24.65 crores. The modernizing and strengthening of Indian recession had made India an interesting investment harbor. So, the past economy present that the real estate rate cannot be automatically increased. Prices of real estate housing property plot were associated with commercial circumstances of state. Even though, we did not have enough appropriate standard regulated approach to measure real estate housing property plot values. Normally, property rate increases along with time and the examined rate should be determined. This examine rate is needed during the sale of housing plot and for the marketability of the house and while applying for the loan. These examined rates are calculated through professional authority. However, defect of this method is that the authority can be biased because of bequeathing interest from sellers, buyers as well as mortgages. So, there is a need for an automatic prediction machine learning model which will help for the prediction of property rates without bias. This machine learning model could advise the less experienced customers along with first-time buyers to interpret whether property values are underrated or overrated. Today, property valuation relies upon different parameters for the housing society as well as for the economy. So, past scrutiny results that house rates are actively dependent on the area of the house along with the geographical location. Then we have tested the parameters to various machine learning model. We have predicted rates of the house using support vector regression machine learning model and linear regression machine learning model and also compared with their actual output values.

In this research paper, we aim to predict housing price rates using these models i.e. Logistic Regression, Decision Tree and Support vector regression with its analogous accuracy and evaluate those based on different types of the error metrics like Root Mean Squared Error(RMSE), R- Squared, Mean Squared Error(MSE) and Mean Absolute Error(MAE).

2. RELATED WORK

Investment is a kind of business activity in which many of them show keen in this modern era. Various objects are often used for investment such as gold, property, and stocks. Property investment has increased significantly. The rise in demand, the uncertainty of economy forces various researchers to find a model that can predict property prices without showing obstacles. That is the reason why Forecasting property value has become a vital field.

From [1] Author scraped data set from duProprio.com and Centris.ca. It consists of 25000 rows and 130 columns. Approximately 70 columns of 130 were scraped. Rest 60 are based on the location of the premises. The author then implemented PCA (Principal Component Analysis) for reducing dimensions. The four techniques used by the author were Support Vector Machine, Linear Regression, K Nearest Neighbors (KNN) and Random Forest Regression and an approach to combine KNN and Random Forest Regression.

Some of the researchers [2] used classifiers for predicting estate values. They took data from Metropolitan Regional Information Systems database. In this, the Author has extracted 15000 records from three sources which contain 76 variables. They wanted to know whether the closing price was lower or higher than the listing price [2]. Models used were. C4.5, RIPPER, AdaBoost, Naive Bayesian.

An author in [3] describes software which is used in the evaluation of real estate prices. This software analyses several real estates, websites of those companies, records their present links to estate's purchase. Data were retrieved from the Czech Republic. This data is retrieved and for every month changes were recorded occurring in real estates. Authors Software can gather 110000 entries per month. Available data was from 2007 to 2015. These entries contain various texts, advertisements, images of properties. Unstructured data which is collected, transformed into a structured form. The data set is further evaluated. Fresh entries which are entered every month is then compared with older entries. In the last phase, clean data is further evaluated and thereby producing several visualizations as per the requirements.

In article [5] using a linear regression technique, the authors predicted the stock market prices. Authors have collected data from the TCS stock Database. RBF and polynomial regression techniques were used by the author along with linear regression and found that latter better than these remaining techniques.

There are various factors which can affect house price rates. In his Rahadi, et al. [3]which segregates into three groups, they are location, concept and Physical condition. Physical conditions refer to the properties which are possessed by house that can be observed by senses, includes house size, total bedrooms, kitchen and garage, garden, land and building area, and house-age. [4]

3. METHODOLOGY

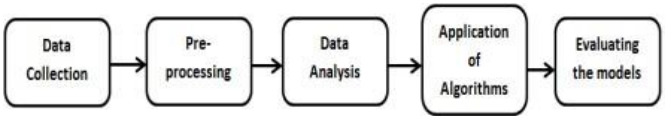

Fig. 1: System Architecture

3.1 Data Collection

The used data set contains 2000 rows with 40 columns that may affect estate prices. Out of 40 columns, only 17 were chosen which really affect the price of houses. The area in sq. meters, Location, Year in which house was built, Total BHK, Garage area, swimming pool area, house selling year and selling price of that house. Here dependent variable is S.P. while others are independent variables. Some of the parameters were numeric form and some of them were in a rating form. Rating form was transformed into numeric form.

3.2 Data preprocessing

Process of transformation of complex data into systematic knowledge is called as Data preprocessing. It finds the missing and unnecessary values in the dataset. The whole dataset is searched for NaN and where the row which consists of NaN will get deleted. This process makes dataset uniform. But our data set contains no NaN data.

Parameters	Description	Datatype
YearBuilt	Original construction date	Numerical
OverallQual	Rates the overall material and finish of the house	Numerical
GrLivArea	Above grade (ground) living area square feet	Numerical
TotalBsmtSF	Total square feet of basement area	Numerical
GarageCars	Size of garage in car capacity	Numerical
GarageArea	Size of garage in square feet	Numerical
FullBath	Full bathrooms above grade	Numerical
WoodDeckSF	Wood deck area in square feet	Numerical
YrSold	Year Sold (YYYY)	Numerical
SalePrice (Dependent Variable)	Selling Price of the house	Numerical

Fig. 2: The Parameters

3.3 Working of various models

The algorithms used by us are Logistic Regression, KNN (K-Nearest Neighbors), Support Vector Machine and Decision Tree. Algorithms implementation is done with the help of SciKit-learn Library present in python. The output is saved in a comma-separated value (CSV) file.

3.3.1 Logistic Regression: It is a binary classification algorithm used when the response variable is dichotomous (1 or 0). Let p indicate probability when Y=1 at X=x, then the model for probability will be:

$$p \equiv P_i(Y = 1|X = x) \equiv \beta_0 + \beta_1 x \qquad (1)$$

Considering p denotes probability, then p should they're between 0 & 1. Therefore, the linear function stated in Equation (1) was unbounded and thus can't further be used for a probability model. In Fig.3, the output is either 1 or 0. Thus a regression lines not be able to build such a classifier model which gives categorical output. This shows that linear regression is not suitable for those class of classification problems where we need to know if the probability output belongs to a particular category. [6]

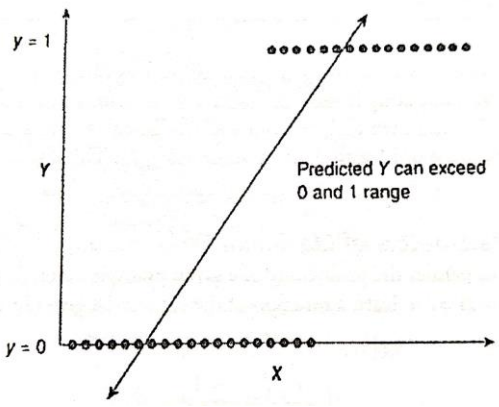

Fig. 3: Graph showing linear regression is not suitable for binary output

Building Logistic Regression Model (Logit Function): Transforming probability to odds eliminate upper bound. If we then take the logarithm of the odds, we also remove the lower bound. Thus, we get the logistic model frame using equation (2).

$$\log[\frac{p_i}{1-p_i}] \equiv \alpha + \beta_1 x_1 + \beta_2 x_2 + ..\beta_i x_i \quad (2)$$

The expression $\log[\frac{Pi}{1-Pi}]$ is called the logit function.

We can solve the logit function for p, to obtain expression given in Eq. (3)

$$p_i \equiv \frac{exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ...\beta_i x_i)}{1 + exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ...\beta_i x_i)} \quad (3)$$

The equation can be simplified by dividing both numerator and denominator by the numerator itself, to achieve the expression for p, given in Eq(4).

$$p_{i1} \equiv \frac{1}{1 + \exp(-\alpha - \beta_1 x_1 - \beta_{21} x_2 + ... \beta_i x_i)} \quad (4)$$

3.3.2 Decision tree: Decision tree algorithms come under the category of supervised learning algorithms. We have built a tree in which the leaf nodes contain the output category. We can predict the class of the output based on the rules generated from the tree structure. Learned trees can be represented as a set of IF-THEN rules as well. The topmost node in a Decision Tree is the root node. A Decision Tree learning algorithm is a variation of the top-down Greedy search algorithm. Two basic algorithms are the Iterative Dichotomiser 3 (ID3) algorithm and the C4.5 algorithm. Those three famous attribute selection measures are gain ratio, information gain and entropy function [6].

Information gain is part of the effectiveness of an attribute in the classification of training data. Entropy is an entity that controls the split in data. It is expected that there will be a reduction in entropy. It computes the homogeneity of examples. The formula for entropy is

$$Entropy(S) \equiv -p_i \log_2 p_i - p_i \log_2 p_i \quad (5)$$

where p stands for the probability of various instances under consideration. [6] The entropy was 0 assuming that all group of S belongs to an equal class and 1 if there are equal numbers of positive and negative examples. [6] It is computed to the base 2 because encoding length is measured in bits. The formula for Information gain is

$$Gain(S,A) \equiv Entropy(S) - \sum \frac{|S|}{|S|} Entropy(S) \quad (6)$$

$$SplitInfo_A(D) \equiv -\sum_{i\equiv 1}^{v} \frac{|D_i|}{|D|} \log_2(\frac{|D_i|}{|D|}) \quad (7)$$

$$GainRatio \equiv \frac{Gain(A)}{SplitInfo_A(D)} \quad (8)$$

This function is the nonlinear logistic regression function. [6]

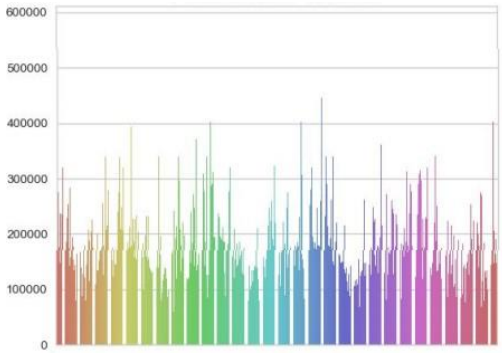

Fig. 4: Predicted price by Logistic Regression

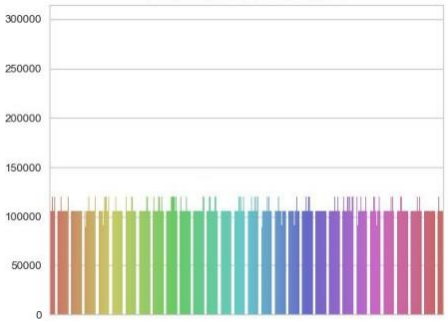

Fig. 5: Predicted price by Decision tree

3.3.3 Support Vector Machine: The SVM are supervised machine learning algorithms that sort data into two categories. Experimentally, it was proved that SVMs have low error rates. (1.1% test error rate for SVMs.) SVMs are today regarded as a crucial illustration of "kernel methods", one of the fundamental fields in machine learning. SVM could be employed for both regression and classification purposes. SVM tries to map an input space into an output space using a nonlinear mapping function Φ such that, the problem of data points become literally separable in the output space. When the points become linearly separable then SVM discovers the optimal separating hyperplane. The SVM automatically discover the optimal separating hyperplane (that is map behind the inside of the input area via Φ^{-1}), which is nothing but a complex decision surface. The objective of the SVM wants to discover optimal separating hyperplane that maximizes the margin of training data. SVM requires training data for finding the optimal separating hyperplane.

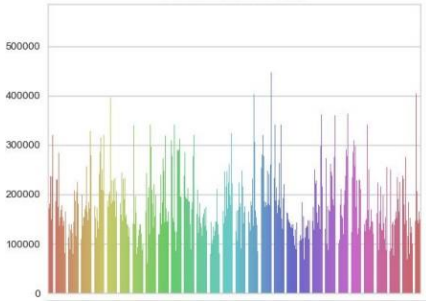

Fig. 6: Predicted price by Support Vector Regression

$$minimize(\frac{1}{2}||w||^2 + C\sum_{m\equiv 1}^{M} + \gamma_m + \gamma_m^*) \quad (9)$$

4. RESULTS

	Accuracy	R-Square	RMSE	MSE	MAE
LR	72.79%	0.978	8922	79604145	6118
DT	84.59%	0.98	217	47184.93	5.68
Svm	66.81%	0.958	14101	1.99E+08	76429

Fig. 7: Results

In the above results, we found that the Decision Tree provides less error values and larger accuracy and R-Square values.

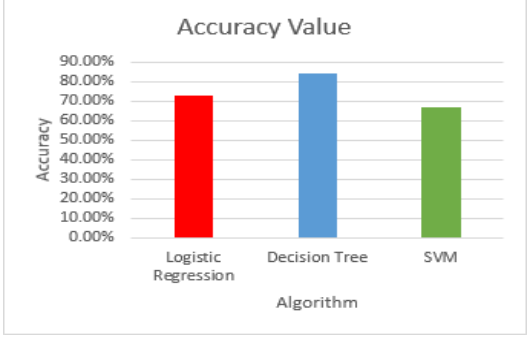

Fig. 8: Accuracy Values

On analyzing, a different model we found Decision Tree performs excellently with topmost accuracy 84.59% and SVM works with least accuracy 66.81%.

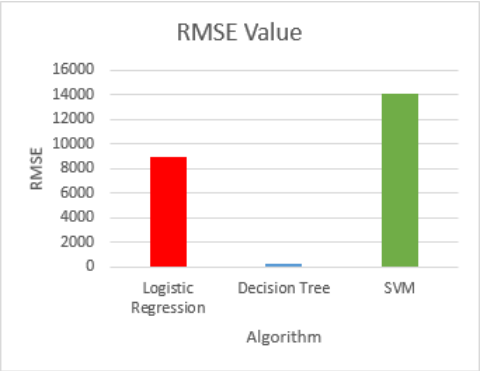

Fig. 9: RMSE Values

Hence, we can say Decision Tree satisfies the dataset and provides an absolutely better accuracy still there is a problem due to the consideration of different noises involve around.

5. CONCLUSION

In this paper, we implemented various machine learning classifiers to know which model is best for a particular dataset. We have calculated the performance of every model and then compared by using different datasets. By running these several models, Decision Tree gave the highest accuracy of 84.59%. Also, Logistic Regression and Support Vector Machines gave 72.79% and 66.81% respectively. From this, we can conclude that the Decision tree overfits the dataset and it gives the best possible accuracy of 84.59%. In this research paper, we have considered a few algorithms which are classifiers. However, the need is that we should train various classifiers, analyze their prediction behavior for a continuous set of values as well. By improvisation of errors, this work could be used for various cities.

6. REFERENCES

[1] Pow, Nissan, Emil Janulewicz, and L. Liu. "Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal." (2014)

[2] 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

[3] 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. 169188, 2015.

[4] V. Limsombunchai, House price prediction: Hedonic price model vs. artificial neural network, Am. J. ..., 2004

[5] Bharatiya, Dinesh, et al. "Stock market prediction using linear regression." Electronics, Communication, and Aerospace Technology (ICECA), 2017 International conference of. Vol. 2. IEEE, 2017.

[6] Vincy Joseph, Anuradha Srinivasaraghavan- "Machine Learning".

[7] Trevor Hastie, Robert Tibshirani, Jerome Friedman- "The Elements of Statistical Learning".

[8] Tom M Mitchell- "Machine Learning"

[9] Saleh Hyatt- "Machine Learning Fundamentals".

