International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 08 | August -2020 ISSN: 2582-3930

Prediction and Impact of Real Estate Property on Economic Using Machine Learning Techniques

Pushpa S[1], Anitha B C[2], Geethika[3], Varshitha[4]

Department of Computer Science,
Cambridge Institute of Technology, Bangalore 560036, India.

Abstract -Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. Housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. Here, the machine learning algorithms helps us to make understand the prediction of the real estate property prices and how it affects the economic growth. In this paper, we will discuss about the prediction of housing prices using machine learning techniques like Naïve Bayes, AdaBoost and Neural Network Algorithm and also discuss about the accuracy of this algorithms. We will also exhibit the various error metrics such as RMSE, MSE, MAE and R-Squared of each models. The motive of this paper is to help the seller to estimate the selling cost of a house perfectly and to help people to predict the exact time slap to accumulate a house.

***Key Words*:**Real estate property, Naïve Bayes, Adaboost, Neural network, RMSE, MSE, MAE, R-Squared.

1.INTRODUCTION

Real Estate property plays a major role in everybody's life. Real estate is known as one of the most important sectors of the economy. They contribute to balancing the economy of a country in as much as it boosts the income of people. The need for a place to stay is going higher and higher as time goes by due to the increasing population, thus many will buy. Real estate enables the economy as it does only speak about houses, but it speaks about almost everything that you see. It gives the space for businesses to operate and it gives off a reputation that a country is doing well. Real estates are forerunners when it comes to giving jobs to many people. From construction to utilization, real estate provides job to many. Such jobs during construction include the engineers, architects, laborers while jobs available after construction greatly depends on how the real estate will be utilized. This is probably one of the main reasons why real estates exists. They are built to bring home and comfort to everyone, especially to families.

Prices of real estate property are related to the economic conditions of the state.Generally, the property values increase with respect to time and decrease with respect to time. Rising house prices, generallyencourage consumer spending and lead to economic growth due to the wealtheffect. A sharpdrop in house prices adversely affects consumer confidence, construction and leads to lower economic growth. A rise in house prices enables homeowners to take out a bigger mortgage. Banks can lead more on the basis of the increased price of the house. Households could use this bigger loan to spend on other items. This can create a significant increase in consumer spending. When there is a fall in house prices, there tends to be a negative wealth effect and a negative impact on economic because households see a fall in house prices, their main form of wealth declines, this reduces their confidence to spend. They are more likely to devote a higher % of their income to try to pay off their mortgage early.[1]

Economic Issues

The USA economic issues were a greater depression in a year 2008.In a year 2000-2006 Indian Real Estate industry was passing through a golden era. It was expected to rise to $90 million by 2015 with a return on investment of 13-16 yields. During 2008-2009, there was a staggering fall in house prices with buyers wary of investing due to further decline in value.The stock market crash of 2008 was the biggest single-day in history due to many people had taken on loans they couldn't afford. Lenders relaxed their strict lending standards to extend credit to people who were less than qualified. This drove up housing prices to levels that many could not otherwise afford. In September 2008, investment firm Lehman Brothers collapsed because of its overexposure to subprime mortgages. It was the largest bankruptcy filing in U.S. history up to that point. The housing slow down caused home prices to decline. Homeowners found themselves "upside down" on their mortgage, meaning they owned more than their home was worth. Faced with job losses and increasing mortgage payments, many lost their homes to foreclosure. Between late 2007 and mid-2009, the period widely referred to as the "Great Depression," the economy lost nearly 8.7 million jobs.[2]

Considering the facts and the studies, it is evident that this growth recession is just a phase and real estate sectors would continue to rise in value. After the 2008-2009 crisis, the years from 2010 have depicted a better picture of the economic situation. Between 2010 and 2012, the economy slowly recovered and thus, the financial situation of both households and real estate companies improved. That development led to higher purchase power and ability to invest, so the demand for real estate increased.

Real Estate Property 2020

At present year 2020 during month of January and February, real estate property sales across different cities were moderate. But during the month ofMarch, the real estate property came to a sudden halt because of nationwidelockdown due tocoronavirus pandemic. In fact, demand for ready homes kept the market afloat in the first quarter of the year until the global crisis in the form of coronavirus struck in India. In the first two weeks of March, property enquiries dipped by over

20% in northern metros and by 7% to 8% in southern metros. Eventually, new project launches were deferred indefinitely and deal closures reduced to almost zero. Owners, too, exited the market due to uncertainties looming the opening up of markets and life getting back to normalcy.

The updated data for housing market predictions from various sources like Realtor.com shows that sales of homes will decline by 15% in 2020.The home prices would flatten out.That's compared to original housing market forecast of a decline of 1.8% in home sales. Single-family housing starts, which were expected to increase by 10% in 2020, are now predicted to decline by 11%. That's mainly due to vigorous social distancing norms and economic uncertainty has compounded this temporary restraint on real estate transactions. According to their statistics, the new listings have declined across the nation's largest metros as sellers wait out the crisis. The positive forecast is that there is expected a short-term bump in sales for late summer and early fall due to pent up buyer demand, fear of the pandemic reducing, and low mortgage rates. Hence, in this case prediction of house prices becomes more important for future.

2. RELATED WORK

There are two important challenges in house price prediction. The first challenge is to identify the number of features that will help to accurately predict the house prices. Now, Property prices depend on various parameters in the economy and society.However, previous analyses show that house prices are strongly dependent on the size of the house and itsgeographical location.[3,4]The usesof various intrinsic parameters (such as number of bedrooms, living area and construction material) was considered for prediction of house prices.[5,6] Then these parameter values were applied to two different machine learning algorithm like linear regression model and Support vector model to predict the price value of the house and compared their output.

The second important challenge that is faced is to find out the machine learning technique thatwill be the most effective when it comes to accurately predicting the house prices.The introduction of resale price prediction of the house using different classification algorithms like Logistic Regression,Decision tree,Naïve Bayes and Random forest and AdaBoost algorithm for boosting upthe weak learners to strong learners. Several factors that are affecting the house resale price includes the physical attributes, location as well as several economic factors persuading at that time.Here, the accuracy based on performance metrics for different datasets and these algorithms are applied and compared to discover the most appropriate method that can be used fordetermining the resale price by the sellers. Hence, it concluded that AdaBoost algorithm produced the best accuracy and logistic regression produced the worst accuracy.[7]The prediction of real estate using hedonic price model and ANN model was also important research paper. Here, the Hedonic price models are basically used to calculate the price of any commodity thatare dependent on internal characteristics as well as external characteristics. The hedonic model basically involves regression technique that considers various parameters such as area of the property, age, number of bedrooms and so on. The Neural Network is trained initially and the weights and biases of the edges and nodes respectively are considered using trial and error method. Training the Neural Network model is a black box method. However, the RSquared value for Neural Network model was greater compared to hedonic model and the RMSE value of Neural Network model was relatively lower. Hence it is concludedthat Artificial Neural Network performs superior than Hedonic model.[8]

The prediction of the house price using linear regression, Lasso Regression, Decision Tree and support vector Regression was calculated. This approach has considered housing data of 3000 properties.Logistic Regression, SVM, Lasso Regression and Decision Tree show the R-squared value of0.98, 0.96,0.81and 0.99 respectively.[9]

3.METHODOLOGY

1. **Gathering of the Data:** The first step for any kind of machine learning analysis is gathering the data, which must be valid. The dataset was taken from one of the largest real-estate portals in the USA. Since the real estate market in the USA is strictly regulated.
2. **Analysing the Data:** Once the gathering of the data has been completed, the next step is to analyse the gathered data. During this phase, the measurable data which affects the house prices were considered. Of course, there were many more parameters that matter as well, such as house condition and location. But these parameters are more subjective and almost impossible to measure, so there were ignored.

Table-1: The Parameters

Parameters	Description	Datatype
Date	Date of Construction	Numerical
Price	Selling price of house	Numerical
Sqft_living	Square feet of the living area	Numerical
Sqft_lot	Total square feet of the land	Numerical
Floors	Floor area of the house	Numerical
Bedrooms	Number of bedrooms	Numerical
Bathrooms	Number of bathrooms	Numerical
Yr_built	Year of building the house	Numerical
Yr_renovated	Year of renovating the house	Numerical
Sqft_basement	Basement of the house	Numerical

3. **Check the correlationbetween parameters**: It is an important need to check for strong correlations among given parameters. If there are, then we should remove one of the parameters. In our dataset there were no strong correlations among values.

4. **Remove outliers from the dataset:** Outliers are observation points that are distant from other observations. For example, in our dataset there was one house with an area of 50 square meters for a price of $500k. Such houses may exist on the market for various reasons, but they are not statistically meaningful.
5. **Algorithms Used:** Once the outliers are removed, we can apply an appropriate machine learning model that fits our dataset. We have selected three algorithms to predict the house price model. The three algorithms that we have selected basically are Ada Boost, Naïve Bayes and the Neural Network algorithms.

4. IMPLEMENTATION OF ALGORITHMS

1. Ada Boost Algorithm

AdaBoost is short for Adaptive Boosting and it is the very popular boosting technique which combines multiple "weak classifiers" into a single "strong classifier". AdaBoost is an ensemble method that constructs a classifier in an iterative fashion. In each iteration, it calls a simple learning algorithm (the weak learner) that returns a classification.The final classification will be decided by a weighted "vote" of the weak classifiers, where each weight is proportional to the correctness of the corresponding weak classifier.

In this algorithm, weak classifiers are selected iteratively from a number of candidate weak classifiers and combined linearly to form a strong classifier for classifying the network data. Let the H= $\{h_f\}$be the set of constructed weak classifiers. The set of training sample data be $(x_1, y_1), \ldots (x_i, y_i), \ldots \ldots (x_n, y_n)$where x_i denotes the $i^{'th}$feature vector, $y_i \in \{+1, -1\}$ is the label of the $i^{'th}$feature vector, denoting whether the feature vector represents a normal behaviour or not; and n is the size of the data set.

This algorithm runs for *T* rounds. Each sample *i* is assigned a weight $w_i(t)$ at any round of *t*. Initially, all weights are set equally, but during the execution of the algorithm these weights are redistributed in order to manipulate the selection process. In each and every round *t* the performance of each single weak classifier assessed. The performances are measured by the weighted error defined as

$$\varepsilon_j = \sum_{i=1}^{n} w_i(t) I[y_i \neq h_j(x_i)$$

Where

$$I_{[y]} = \begin{cases} 1, & y = true \\ 0, & y = false \end{cases} \qquad (1)$$

At the end of each round, the classifier *h* with the lowest error rate of ε_t based on the equation 1 is selected and stored as best classifier h_t of round *t*. Then a confidence α_t is computed as

$$\alpha_t = \frac{1}{2} log(\frac{1-\varepsilon_t}{\varepsilon_t})$$

It can be interpreted as the quality of hypothesis h_t; the lower error rate ε_t, the higher confidence α_t. Intuitively, $\alpha_t$measures the importance that is assigned to h_t. Note that $\alpha_t > 0$ if $\varepsilon_t > \frac{1}{2}$and that $\alpha_t$gets larger as $\varepsilon_t$gets smaller. Finally, the weights are redistributed and normalized as follows

$$w_i(t+1) = (\frac{w_i(t)\exp(-\alpha_t y_i h_t(x_i))}{z_t}$$

By increasing the weights of samples that were misclassified by $h_t$favours, in the next round, these difficult samples are handled correctly. The effect of this rule is to increase the weight of examples misclassified by h_t and to decrease the weight of correctly classified examples. Thus, the weight tends to concentrate on "hard" examples. z_t denotes the normalization factor which ensures the sum of all weights to be +1. The normalization factor is defined as

$$z_t = \sum_{k=1}^{n} \exp(-\alpha_t y_i h_t(x_k))$$

The final hypothesis *H* is a weighted majority vote of the *T* selected weak classifiers where α_t is the weight assigned to hypothesis h_t

$$H(x) = \sin(\sum_{t=1}^{T} \alpha_t h_t(x))$$

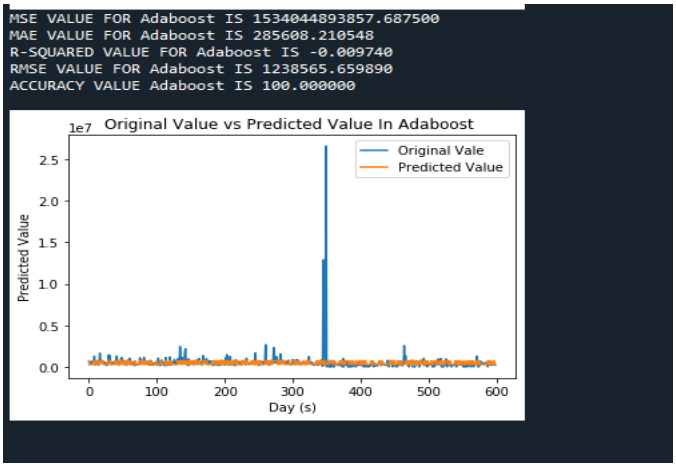

Fig-1: Performance of Ada Boost Model

Figure 1 shows the performance of Ada Boost algorithm with 100% accuracy. The Ada Boost algorithm is the best model among other models. The RMSE value was approximately less when compared to Naïve Bayes.

2. Neural Network

Neural networks are artificial systems that were inspired by biological neural networks. Neural networks are based on computational models for threshold logic. Threshold logic is a combination of algorithms and mathematics. Neural networks are based either on the study of the brain or on the application of neural networks to artificial intelligence.

Multi-layer Neural Network

A Multi-Layer Perceptron (MLP) or Multi-Layer Neural Network contains one or more hidden layers (apart from one input and one output layer). A multi-layer perceptron can learn nonlinear functions.

This neuron takes as input $x_1, x_2, \ldots, x_3$ (and a +1 bias term), and outputs f(summed inputs + bias), where f(.) called the activation function. The main function of Bias is to provide every node with a trainable constant value (in addition to the normal inputs that the node receives). Every activation function (or non-linearity) takes a single number and performs a certain fixed mathematicaloperation on it. There are several activation functions as

Sigmoid Functiontakes real-valued input and squashes it to range between 0 and 1 as

$$\sigma(x) = \frac{1}{(1 + \exp(-x))}$$

Tanh Functiontakes real-valued input and squashes it to the range [-1, 1] as

$$\tanh(x) = 2\sigma(2x) - 1$$

ReLu stands for Rectified Linear Units. It takes real-valued input and thresholds it to 0 (replaces negative values to 0) as

$$f(x) = \max(0, x)$$

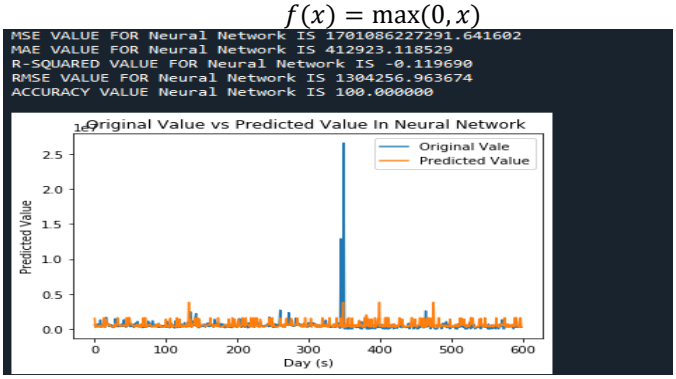

Fig-2: Performance of Neural Network Model

Figure 2 shows the Neural network is second best model. This model shows less RMSE when compared to other two algorithms.

3. Naïve Bayes

The Naive Bayes Classifier is based on the Bayes' theorem. The Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation

$$P(y|X) = \frac{P(X|y)\,P(y)}{P(X)}$$

Now, if any two events A and B are independent, then P (A, B) = P(A)P(B).

Hence, we reach to the result

$$P(y|x_{1.....x_n}) = \frac{P(x_1|y)P(x_2|y).....P(x_n|y)P(y)}{P(x_1)P(x_2).....P(x_n)}$$

which can be expressed as

$$P(y|x_{1.....x_n}) = \frac{P(y)\ \prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2).....P(x_n)}$$

Now, as the denominator remains constant for a given input, we can remove that term

$$P(y|x_{1.....x_n}) \propto P(y)\ \prod_{i=1}^{n} P(x_i|y)$$

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible valuesof the class variable *y* and pick up the output with maximum probability. This can be mathematically expressed as

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

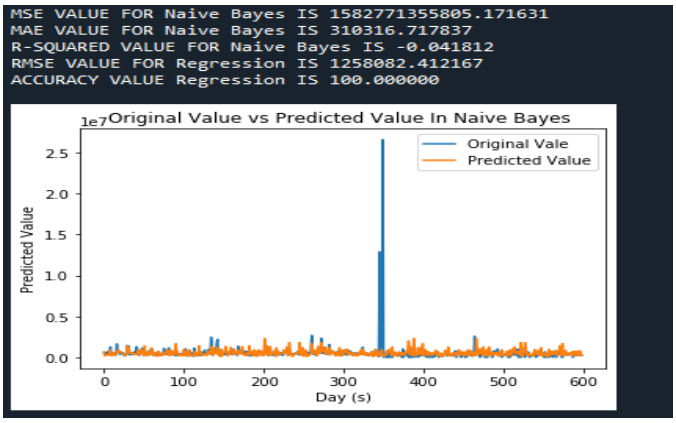

Fig-3: Performance of Naïve Bayes Model

Figure 3 shows the performance of naïve Bayes and its error. As we can observe that RMSE of the naïve Bayes model shows more value than Ada Boost model.

3. RESULT &CONCLUSIONS

This paper mainly concentrates on how real estate property affects the economic growth and how it reflects on people's career. As we know that housing prices keep changing day in and day out and sometimes are hyped rather than being based on valuation. A sharp drop in house prices adversely affects consumer confidence, construction and leads to lower economic growth. Even rise in real estate property also affects the economic. As we also discuss about great depression in year 2008-2009 which lead to economic crises. At present year 2020, we also saw sales of real estate growing moderately during the month of January and February but during the March month because of lockdown due to coronaviruses pandemic the real estate prices once again fall down. In this research paper, we have used machine learning algorithms to predict the real estate prices. We also computed the performance of the algorithms along with the step wise implementation of algorithms. From the above computation of the algorithms, we found that Ada Boost is first best model with respect to accuracy and Neural network the second best model.

ACKNOWLEDGEMENT

This work was supported by the Cambridge Institute of Technology. We thank our project guide, Prof. Pankaja K who gave comments and suggestions, which have helped us to improve the paper substantially.

REFERENCES

1. Tejvan Pettinger: How the housing market affects the economy. https://www.economicshelp.org/blog/21636/housing.
2. The Market Crash of 2008. https://www.wealthsimple.com/en-ca/learn/2008-Market-crash.
3. D. Belsley, E. Kuh, Welsch.: RegressionDiagnostics: Identifying Influential Data and Source of Collinearity. NewYork: John Wiley, 1980.
4. J. R. Quinlan, "Combining instance-based and the model-basedlearning," Morgan Kaufmann, 1993, pp. 236–243.
5. S. C. Bourassa, E. Cantoni, and M. Hoesli, "Predicting house prices with spatial dependence: a comparison of alternative methods," Journal of Real Estate Research, vol. 32, no. 2, pp. 139–160, 2010.
6. S. C. Bourassa, E. Cantoni, and M. E. Hoesli, "Spatial dependence, housing submarkets and house price prediction," eng, 330; 332/658, 2007, ID: unige:5737.
7. P. Durganjali, M. Vani Pujitha, "House Resale Price Prediction Using Classification Algorithms," IEEE 6th International Conference on smart structures and systems ICSSS 2019.
8. Limsombunchai, Visit. "House price prediction: hedonic price model vs. artificial neural network. "New Zealand Agricultural and Resource Economics Society Conference.2004.
9. Neelam Shinde, Kiran Gawande, "Valuation of House Prices using Predictive Techniques,"International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835Y.
10. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.