House Price Prediction Using Machine Learning: A Review Paper

Aavish Baraptre[1], Tanveer Chawla[2], *Santosh Bothe*[3]

[1,2]*Undergraduate, Computer Engineering Department, Mukesh Patel School of Technology Management and Engineering (MPSTME), NMIMS, Shirpur, Maharashtra*

[3]*Assistant Professor, Computer Engineering Department*

Mukesh Patel School of Technology Management and Engineering (MPSTME), NMIMS, Shirpur, Maharashtra

ABSTRACT

Real estate is the most obvious industry in our environment. Housing prices are constantly changing from day to day and are sometimes fired rather than based on estimates. Predicting real estate prices by real factors is a key element of our research project. Here we aim to make our test based on all the basic parameters considered while determining the value. The proposed process looked at the more refined aspects used in house price calculations and provided more accurate forecasts. It also provides a brief overview of the various graphic and calculation techniques that will be needed to predict the price of a house. This paper contains how the real estate model works with the help of machine learning and what data is used in our proposed model.

Keywords: *Machine learning, Regression Technique, Classification Technique, Cross validation Technique, K-means*

1. INTRODUCTION

A house / home is a basic need for a person and their prices vary from place to place depending on available resources such as parking, space, etc. The price of a house is a factor that worries a ton of residents whether they have a rich or white collar as one cannot judge or measure the value of a house based on the location or offices available. Buying a home is one of the biggest and most important options for a family as it spends all of their investment money and often includes it under a mortgage. It is a difficult task to predict accurate home prices. Our proposed model will enable us to predict house prices.

1.1 Objective

This project is intended to predict housing prices and achieve better and more accurate results. The stack algorithm is used in various back-to-back algorithms to see which algorithm has the most accurate and precise results. This can be very helpful to people because house prices call for a topic that affects many citizens whether rich or middle class as someone who cannot judge or estimate the price of a home on the basis of available space or resources. To accomplish this task, a python program language is used. Python is a high-level programming language for the general purpose program. Enables clear programs on small and large scales. Easy-to-read language.

1.2 Real Estate Market

The real estate market that connects the economy consists of a number of specific items: regulations, building cycles and price trends, market forces, and sub-markets and assessments. Economists use these items to predict real estate prices accurately. First, regulations make a big issue. Regulations can tell us whether a grant can respond to the change required. In addition, they tell us something about the risk of new developments or recurrences that affect the required return or expected profit marks. [1] And they should protect us and give us higher land prices when we benefit from better planning. Regulations can also address issues of civil and private rights.

1.3 Machine Learning

Machine learning is an Artificial Intelligence field that enables PC structures to learn and improve working with information. It is used to study the construction of algorithms that perform data predictions. Machine learning is used to do many computer tasks. It is also used to make predictions about computer use. Machine learning is sometimes used to build complex models. The basic point of machine learning is to allow PCs to learn things naturally without human help. Machine learning is very useful and widely used around the world. [2] The process of machine learning involves providing information and computer training by building machine learning models with the help of various algorithms. Machine learning can be used to perform various programs such as application detection face, etc. Machine learning is a field of software engineering that has revolutionized the way information is multiplied.

1.4 Programming Language and libraries

In this Review paper the main language we have used is Python as it works well and is easy to use. Enables clear programs on small and large scales. Python reinforces various levels of editing including object programming, usefulness and process. Python is an easy-to-read language. It uses English keywords and some programming languages use punctuation. Python uses white space in contrast to wavy sections to measure squares. Python was developed especially for easy coding. [2] Python supports various libraries such as Pandas, NumPy, SciPy, Matplotlib etc. It supports various packages such as Xlsx Writer and Xl Rd. Python is a very useful language for web development and programming. It is often used to create web applications. It has been well used to read and translate texts. It can be used effectively to make complex science. Python has acquired a language that is well-known for its versatility.

2. PROPOSED MODELS

Our database has a variety of important parameters and data mining has been the root of our system. Initially we cleaned up our entire database and reduced foreign prices. In addition, we rated each parameter according to its importance in determining the pricing of the system and this led us to increase the value held by each parameter in the system. We have compiled a list of different machine learning algorithms and tested our system with various combinations that can ensure the reliability of our results.

2.1 Linear Regression

Linear regression is the easiest way to predict. It uses two elements such as forecasting variables and the most important variables first of all whether the forecast is variable and strategic. These regression values are used to define the relationship between a single variable and one or more independent variables. The regression equation calculation with a single dependent and a single independent variation is defined by the formula. [8]

b = y + x*a

where, b = estimated dependent variable score, y = constant, x = regression coefficient, and a = score on the independent variable.

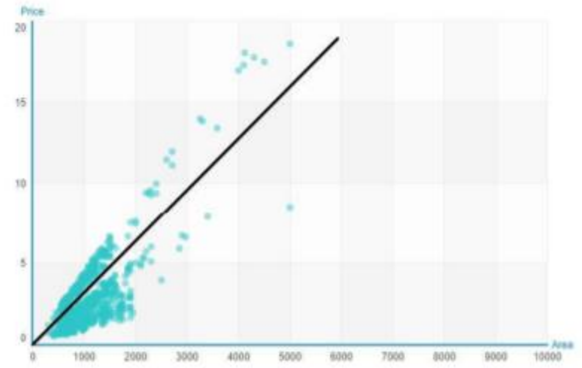

Figure 1 Linear regression scatter plot

2.2Forest Regression

Forest regression uses a process called Bagging of trees. The main idea here is to decorrelate several trees. We then minimize tree variations by measure. [3] Using this method, a large number of decision trees are created.

The random forest training algorithm uses the process of assembling, or bagging, to tree learners. [7] Given a training set X = x1, ..., x_n with responses Y = y1, ..., y_n, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For b = 1, ..., B:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.

2. Train a classification or regression tree fb on X_b, Y_b.

After training, predictions for unseen samples a' can be made by averaging the predictions from all the individual regression trees on a':

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on a':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} (f_b(x') - \hat{f})^2}{B-1}}.$$

2.3 Extreme Gradient Boosting (XGBoost)

XGBoost is an unparalleled machine learning program for tree boosting. The program is available as an open-source pack-age. The program has had a significant impact and is widely known for the various challenges of machine learning and data mining [6]. The most important reason why XGBoost succeeds in its failure in all cases. The system works more than ten times better than popular solutions available on a single machine and scales millions of examples in distributed or limited memory settings. The decline of XGBoost is due to a number of large-scale programs and algorithmic adjustments including a tree-based learning algorithm for managing small data and a standard sketch system that enables weight management in tree learning thinking. Augmented retreat is a type of learning technique that produces prediction with the help of decision-making trees that often incorporates many weak predictor models [10]. This Growth algorithm takes real-life value y and seeks to measure F (x) in the form of an estimated hi (x) scale from class H called weak students:

$$F(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const}.$$

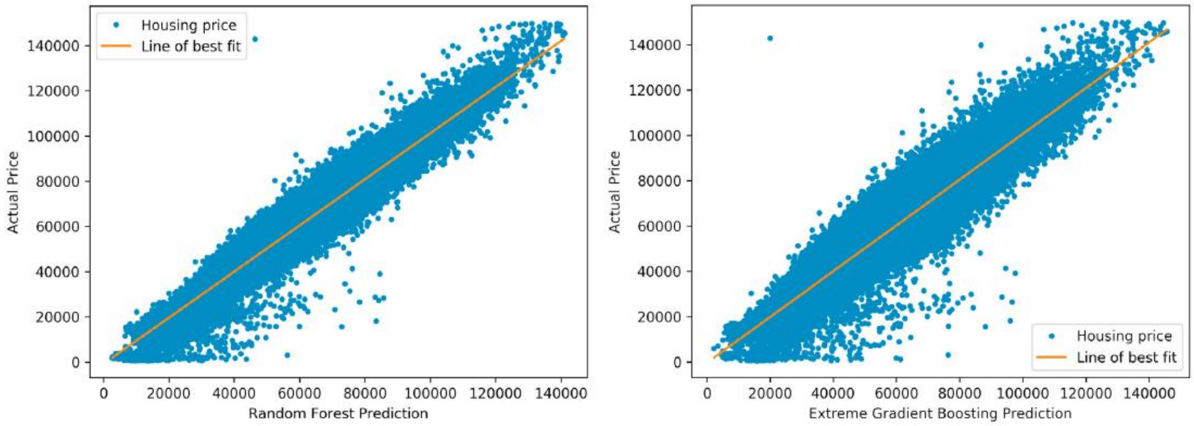

Figure 2 Prediction of house price by Random Forest and XGBoost

2.4 Neural Networks

An artificial network learning algorithm, or neural network, or just a neural net, is an integration learning system that uses a network of tasks to understand and interpret the input of a single form out of demand, often the other way around. The concept of the neural artificial network is inspired by human biology and how the human brain neurons work together to understand inputs from human hearts.

Neural networks are one of the tools and methods used in machine learning algorithms. The neural network itself can be used as a component in many machine learning algorithms to process complex data input in a computer-incomprehensible space.

Machine learning algorithms that use neural networks usually do not need to be programmed with specific rules that define what to expect from inputs. The neural net learning algorithm instead learns by processing the many labeled examples (e.g. data with "answers") provided during training and using this answer key to learn what input features are needed to create a specific result. [4] Once a sufficient number have been processed, the neural network can begin to process new, invisible inputs and successfully restore accurate results. The more examples and variations of the inputs perceived by the system, the more effective the results are because the system learns from experience.

Neural networks can be used for a wide range of problems and can explore many types of input, including photos, videos, files, info, and more. Nor do they need a clear program to translate the content of the input.

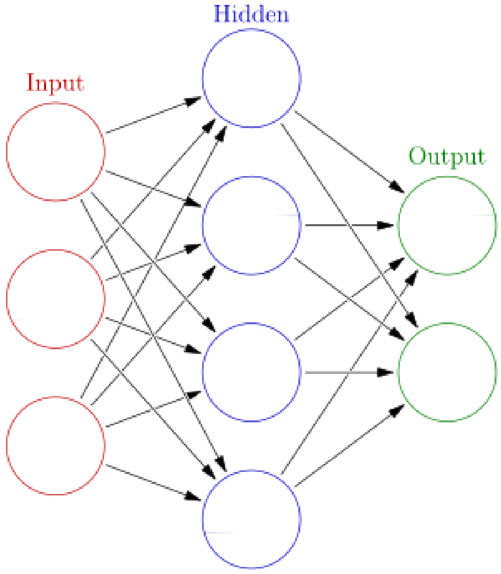

Figure 3 A simple neural network having a single hidden layer

2.5 Procedure

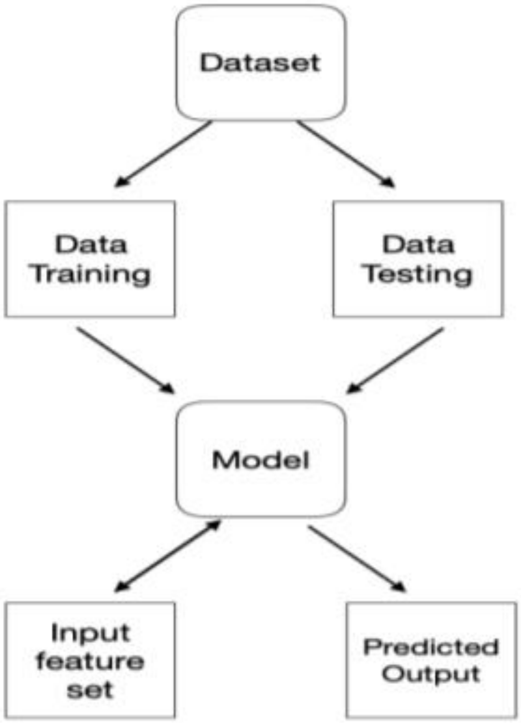

Figure 4 Procedure of training a model.

2.5.1 Data Collection

Data collection is the process of gathering information on variables in a systematic manner. This helps to find answers to many questions, the hypothesis and to evaluate the results. Data collection is a way to a public event and to measure data on objects that are based on a structured framework, at the same time empowering a person to answer relevant questions and evaluate results.

2.5.2 Data Analysis and Visualization

Data Visualization is symbolic or graphical representation of information. It makes it easy to understand complex concepts or identify new patterns. Data visibility is seen in many orders as visual-like visual cuts. Includes the creation and investigation of information displays. To convey information clearly and effectively, the representation of information uses measurable images, sites, data formats and unique tools. Active perception helps customers with distinction and thinking in detail and validation.

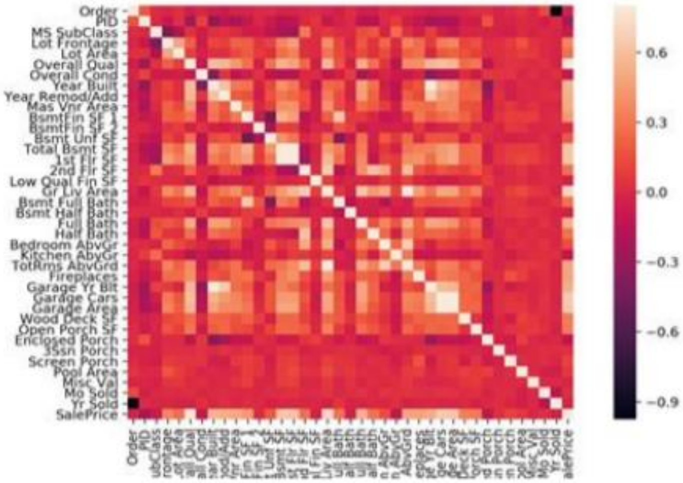

Figure 5 Correlation

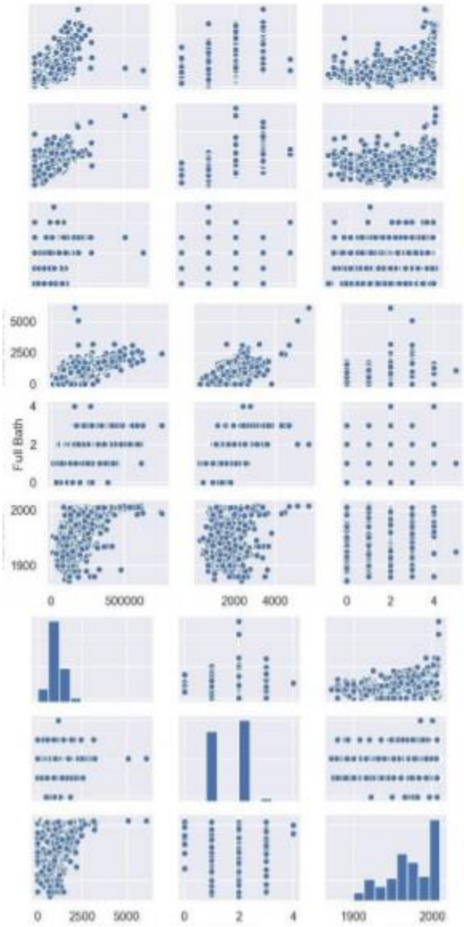

Figure 6 Scatter plot

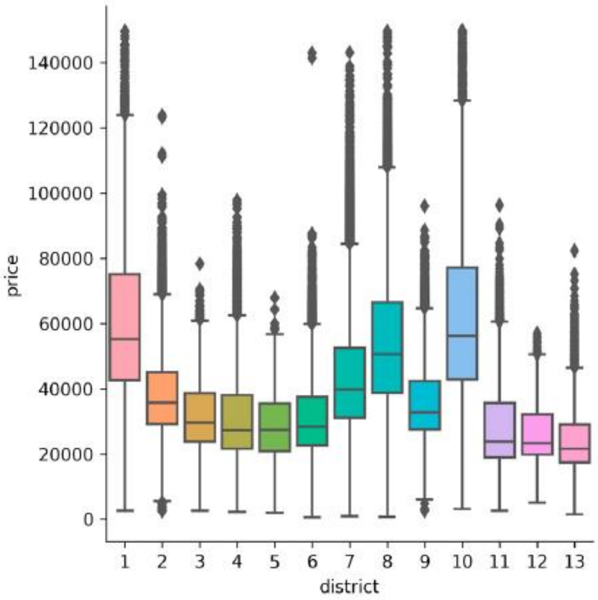

Figure 7 Correlation between District and Price

2.5.3 Data Cleaning

Data cleaning is the process of finding and removing errors to increase the amount of data. Data cleaning is done with the help of data dispute tools. It is a way of identifying and updating basic records from a record set, table or database. It finds missing information and replaces it with obscene information. The details are changed to ensure that they are accurate and correct. Data cleaning is the process of sorting and updating erroneous records from a record set, table or database. It is a way of looking at insufficient details and then inserting dirty information. The details are changed to ensure that they are accurate and correct. Used to make database predictions. The purpose of the data purification policy is to eliminate and dispel errors to create power information limitations.

2.5.4 Model Selection

Prior to modeling, the data must be properly processed so that the models can read the patterns correctly. Specifically, numerical values are assigned equally, while class values were written in a single code. After processing, the database included most features. The diagram below shows the variations in the collected data. Since most of the features are categorized, it is clear that the variability is likely to change in the 30th segment. After that, the database was divided into a set of training and tests set at a ratio of 4: 1 using a scikit-learn package [9]. The test function used for this error is Root Mean squared Logarithmic Error (RMSLE).

This function is illustrated as follow:

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2}$$

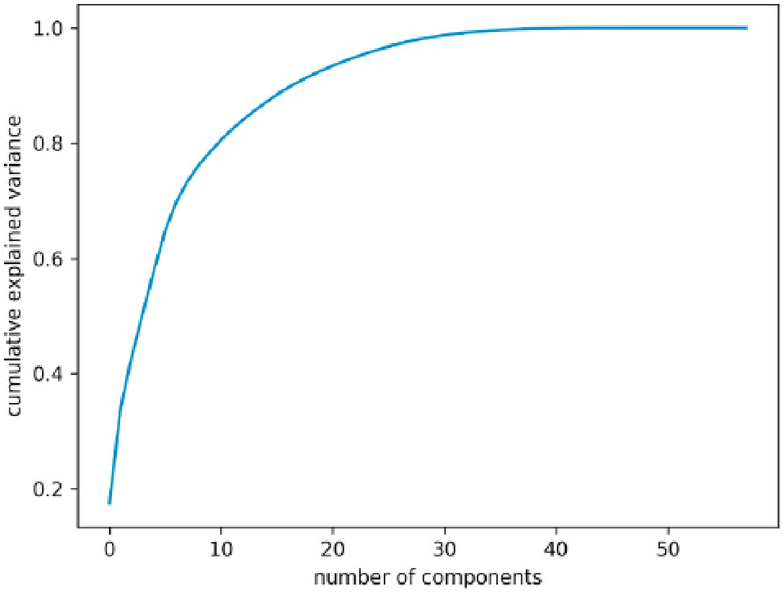

Figure 8 Cumulative Explained Variance

3. RESULT

To achieve results, various data mining techniques are used in the python language. There are a variety of factors that affect the price of a home that are considered and worked on. Machine learning is considered to complete the job you are looking for. First, data collection is done. After that data cleaning is done to remove all errors from the data and make it clean. After that the data processing was done. Thereafter, with the help of data identification, various sites were created, which aimed to reflect data distribution in a variety of ways. In the end, the cost of real estate was determined precisely. This can be achieved because a simple stacking algorithm is used to improve the understanding of the various algorithms used in our housing price database to provide better results.

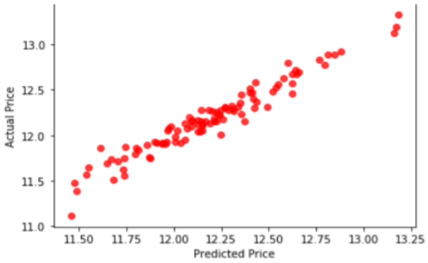

Figure 9 Scatter Plot

4. CONCLUSION

A system designed to provide accurate forecasts for housing prices has been developed. The system makes good use of Linear Regression, Forest regression, Boosted regression. The efficiency of the algorithm has been greatly enhanced by the use of Neural networks. The system will satisfy customers by providing accurate output and preventing the risk of investing in the wrong house. Additional customer benefit features can also be added to the system without interfering with its basic functionality. A major future update could be the addition of major cities to the database, which will allow our users to explore more homes, find more accuracy and thus reach the right decision.

5. FUTURE WORK

System accuracy can be improved. Several quotes can be added to the program if the size and power of the computer increases in the program. In addition, we can integrate a different UI / UX approach to better see results in a more interactive way using the unpopular taxpayer we see. Also, a learning program can be created that will gather users feedback and history so that the system can display the most relevant results for the user according to his or her preferences. In the future, many algorithms can be used in this database such as decision tree, Naïve Bayes, SVM etc. And get their proper details and use it to predict the best result and increase accuracy. The KNN algorithm can also be used to predict accuracy. The k-means algorithm can also be used. With the help of these algorithms, house prices are accurately predicted. Therefore, it can be very helpful for the government and the people themselves. Regression algorithms were initially adopted for our project but in the future, this can be obtained and using phase algorithms.

References

[1] Y. Luo, "Residential Asset Pricing Prediction using Machine Learning," 2019 International Conference on Economic Management and Model Engineering (ICEMME), Malacca, Malaysia, 2019, pp. 193-198, doi: 10.1109/ICEMME49371.2019.00046.

[2] M. Jain, H. Rajput, N. Garg and P. Chawla, "Prediction of House Pricing using Machine Learning with Python," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 570-574, doi: 10.1109/ICESC48915.2020.9155839.

[3] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How many trees in a random forest?” In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7376 LNAI, 2012, pp. 154–168, ISBN: 9783642315367. DOI: 10 . 1007 / 978 - 3 - 642 - 31537-4\ 13

[4] J. Schmidhuber, “Multi-column deep neural networks for image classification,” in Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), ser. CVPR ’12, Washington, DC, USA: IEEE Computer Society, 2012, pp. 3642–3649, ISBN: 978-1-4673-1226-4.

[6] Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16 2016. doi:10.1145/2939672.2939785.

[7] The elements of statistical learning, Trevor Hastie - Random Forest Generation

[8] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

[9] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 2011;12:2825–30

[10] Friedman, J. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29(5):1189–1232