Real Estate Price Prediction

Hemalatha K N[1], Savitha S J[2], Yuktha S[3], Rashmi A [4]
[1]Associate Professor, Dept of Computer Science
[2, 3, 4]Dept of Computer Science
[1, 2, 3, 4] AIT, Bengaluru, India

Abstract- *This provides an overview about how to predict house costs utilizing different regression methods with the assistance of python libraries. The proposed technique considered the more refined aspects used for the calculation of house price and provide the more accurate prediction. It also provides a brief about various graphical and numerical techniques which will be required to predict the price of a house. The real estate market is one of the most competitive in terms of pricing and same tends to vary significantly based on numerous factors; forecasting property price is an important module in decision making for both the buyers and investors in supporting budget allocation, finding property finding stratagems and determining suitable policies hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with accuracy.*

I. INTRODUCTION

House/Home are a basic necessity for a person and their prices vary from location to location based on the facilities available like parking space, locality, etc. The house pricing is a point that worries a ton of residents whether rich or white collar class as one can never judge or gauge the valuing of a house based on area or offices accessible. Buying of a house is one of the greatest and significant choice of a family as it expends the entirety of their investment funds and now and again covers them under loans. It is the difficult task to predict the accurate values of house pricing. Our proposed model would make it possible to predict the exact prices of houses.

A. *Machine Learning:* Machine Learning is a field of Artificial Intelligence which enables PC frameworks to learn and improve in execution with the chine learning is used to perform a lot of computing tasks. It is also used to make predictions with the use of computers. Machine learning is sometimes also used to devise complex models. The principle point of machine learning is to permit the PCs to learn things naturally without the assistance of people. Machine learning is very useful and is widely used around the whole world. The process of machine learning involves providing data and then training the computers by building machine learning models with the help of various algorithms. Machine learning can be used to make various applications such as face detection application, etc. Machine Learning is a field in software engineering that has changed the way of examining information colossally.

B .Python: Python is an elevated level programming language for broadly useful programming. It was created by Guido Van Rossum and released in 1991. It enables clear programming on both small and large scales. Python bolsters various programming standards including object arranged, useful and procedural. Python is an easily readable language. It uses English keywords whereas other programming languages use punctuations. Python utilizes whitespace space as opposed to wavy sections to delimit squares. Python was mainly developed to read codes easily. Python supports various libraries such as Pandas, NumPy, SciPy, Matplotlib etc. It supports various packages such as Xlsx Writer and XlRd. Python is an exceptionally helpful language for web improvement and programming advancement. It tends to be utilized to make web applications. It very well may be utilized to peruse and alter documents. It very well may be used to perform complex science. Python has gotten a very well-known language since it can chip away at various stages. Python code can be executed when it is composed. Python is a very significant language since the program is updated without investing additional exertion and energy. Python bolsters many working frameworks.

II. LITERATURE SURVEY

There are a couple of components that impact house costs. In this exploration, partition these components into three essential get together, there are state of being, thought and territory [1].States of being are properties constrained by a house that can be seen by human recognizes, including the range of the house, the amount of rooms, the availability of kitchen and parking space, the openness of the yard nursery, the zone of land and structures, and the age of the house [2].While the thought is an idea offered by architects who can pull in potential buyers, for instance, the possibility of a moderate home, strong and green condition, and world class condition. Zone is a critical factor in shaping the expense of a house. This is in light of the fact that the zone chooses the normal land cost [3].Besides, the territory furthermore chooses the basic passage to open workplaces, for instance, schools, grounds, crisis facilities and prosperity centers, similarly as family preoccupation workplaces, for instance, strip malls, culinary visits, or much offer awesome landscape [4].Kilpatrick [5] showed the usefulness of time-series regression model which used economic data to provide forecast of Central Business District (CBD) land price in moving market. Wilson etal.[6] studied the residential property market accounts for a substantial proportion of UK economic activity. Valuers estimate property values based on current bid prices. In this paper, the national housing transaction data was trained using Artificial Neural Networks (ANN), which forecasts future trend of the housing market.Mark and John [7] developed a regression model with vacant land sales. The model explained up to 93% of the market values.Wang and Tian [8] used the wavelet Neural Network (NN) to forecast the real estate price index. This kind of wavelet NN integrated the merit of the wavelet analysis and the tradition NN. It also compared the forecasting result with smoothing method and the NN forecast.Zhangming [9] forecasted the real estate price index by using the Back Propagation (BP) NN. The BPN used the sigmoid function.Tinghao [10] used the Auto Regressive Integrated Moving Average (ARIMA) model and carried the demonstrative analysis on year data from 1998 to 2006. He used the established model to make the forecast to the real estate price index of 2007. A hedonic regression on the price of land suggested that de facto policy differences between political jurisdictions have had a significant effect on land prices between1970 and 1980.Steven and Albert [11] used 46,467 residential properties spanning 1999 - 2005 and demonstrated that using matched pairs that relative to linear hedonic pricing models, ANN generate lower dollar pricing errors, had greater pricing precision out- of-sample, and extrapolate better from more volatile pricing environments. ANN is better suited to hedonic models that utilize large numbers of variables.Sampath kumar and Santhi [12] studied the land price trend of Sowcarpet which is the central part. They developed statistical model using economic factors and predicted that the annual rise in land price would be of 17%.Urmila [13] reported that the past trends were analysed to ascertain the rate of growth or decline and the trends are used in forecasting. Economic parameters might be introduced to formulate more realistic relationship.Some of the other techniques they Mansural Bhuiyan and Mohammad Al Hasan 2016 [14] use is regression, deep learning to learn the nature of models from the previous results (the property/land which were sold off previously which are used as training data).There are different models used such as linear model data using only one feature, multivariate model, using several features as its input and polynomial model using the input as cubed or squared and hence calculated the root mean squared error (RMS value) for the model.

III. SYSTEM ARCHITECTURE AND DESCRIPTION

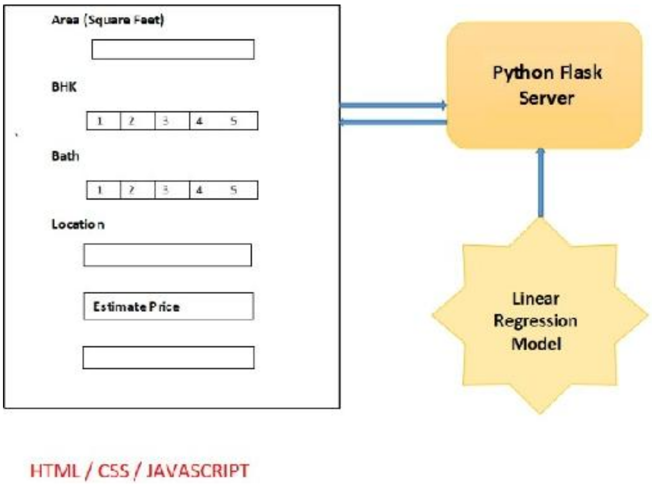

Figure 1. High level block diagram

A. Linear Regression Model: Regression analysis is widely used for forecasting.Regression analysis is used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. If more independent variables are added, it is able to determine an estimating equation that describes the relationship with greater accuracy. Multiple regressions look at each independent variable and test whether it contributes significantly to the way the regression describes the data.

B. *Python Flask Server:* Flask is a small and lightweight Python web framework that provides useful tools and features that make creating web applications in python easier. It gives developers flexibility and is a more accessible framework for new developers since you can build a web application quickly using only a single Python file.

IV. WORKING ANDMETHODOLOGY

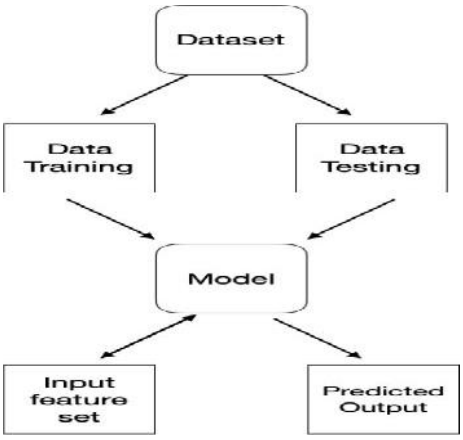

Figure2. Block Diagram

A. *Data Collection :* Data collection is the process of gathering in formation on variables in a systematic manner. This helps in finding answers too many questions, hypothesis and evaluate outcomes. Data collection is the way toward social event and estimating data on focused factors in a built up framework, which at that point empowers one to address pertinent inquiries and assess results. Information assortment is a part of research in all fields of study including physical and sociologies, humanities and business. While strategies differ by discipline, the accentuation on guaranteeing precise and legitimate assortment continues as before. It has been attempted for various datasets on Kaggle, which would suite our project objective. After looking at a lot of datasets, this dataset is found. It is a house pricing dataset in the city of Ames. This dataset is a very popular machine learning dataset with less scope of errors and variations.

B. *Data Visualization :* Data Visualization is the pictorial or graphical representation of information..It enables to grasp difficult concepts or identify new patterns. Data Visualizations seen by numerous orders as a cutting edge likeness visual correspondence. It includes the creation and investigation of the visual portrayal of information. To impart data plainly and effectively, information representation utilizes measurable illustrations, plots, data designs and different apparatuses.. Effective visualization assists customers with separating and reason about data and verification. It makes complex data progressively accessible, reasonable and usable. Customers may have explicit logical endeavors, for instance, making assessments or getting causality, likewise, the structure standard of the reasonable (i.e ., indicating examinations or demonstrating causality) follows the undertaking. Data Visualization is both a craftsmanship and a science. It is viewed as a piece of particular estimations by a couple, yet what's more as a grounded theory improvement device by others. Extended proportions of data made by Web activity and an expanding number of sensors in the earth are suggested as "enormous data" or Web of things. Dealing with, analyzing and passing on this data present good and orderly challenges for data portrayal. The field of data science and experts called data scientists help address this test.

C. *Data Pre-processing :* It is the process of transforming data before feeding it into the algorithm. It is utilized to change over crude information into a clean dataset. It is an information mining strategy that includes moving crude information into a justifiable organization. The result of data preprocessing is the last dataset utilized for preparing and testing reason. Data preprocessing is an information mining procedure which is utilized to change the crude information in a helpful and productive format. In any Machine Learning procedure, Data Preprocessing is that progression wherein the information gets changed, or Encoded, to carry it to such an e xpress, that now the machine can without much of a stretch parse it. Pre-dealing with insinuates the progressions applied to our data before dealing with it to the estimation.Data Pre-processing isa system that is used to change over the rough data into an ideal enlightening assortment. In a manner of speaking, at whatever point the data is amassed from different sources it is assembled in rough setup which isn't feasible for the examination. Genuine in formation for the most part contains clamors, missing qualities, and perhaps in an unusable organization which can't be legitimately utilized for Machine Learning models. Data pre processing is required errands for cleaning the information and making it appropriate for an Machine Learning model which likewise expands the precision and proficiency of a Machine Learning model.

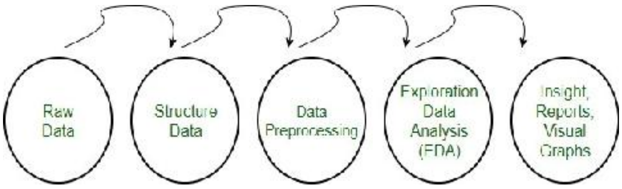
Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs

Figure 3. Data Pre-processing

D. *Data Cleaning :* Data cleaning is the process of detecting and removing errors to increase the value of data. Data cleaning is carried out with the help of data wrangling tools. It is the way toward identifying and amending off base records from a record set, table or database. It finds the deficient information and replaces the messy information. The information is changed to ensure it is exact and right. Information cleaning is the way toward distinguishing and revising mistaken records from a record set, table or database. It is the way toward recognizing inadequate information and afterward supplanting the messy information. The information is changed to ensure that it is exact and right. It is utilized to make a dataset predictable. The principle objective of data cleaning is to distinguish and expel blunders to build the estimation of information in dynamic.

V. CONCLUSION

The sales price for the houses are calculated using different algorithms. The sales prices have been calculated with better accuracy and precision. This would be of great help for the people. To achieve these results, various data mining techniques are utilized in python language. The various factors which affect the house pricing should be considered and work upon them. Machine learning has assisted to complete out task. Firstly, the data collection is performed. Then data cleaning is carried out to remove all the errors from the data and make it clean. Then the data pre-processing is done. Then with help of data visualization, different plots are created. This has depicted the distribution of data in different forms. Further, the preparation and testing of the model are performed. It has been found that some of the classification algorithms were applied on our dataset while some were not. So, those algorithms which were not being applied on our house pricing dataset are dropped and tried to improve the accuracy and precision of those algorithms which were being applied on our house pricing dataset. To improve the accuracy of our classification algorithms, a separate stacking algorithm is proposed. It is extremely important to improve the accuracy and precision of the algorithms in order to achieve better results. If the results are not accurate then they would be of no help to the people in predicting the sales prices of houses. It also made use of data visualization to achieve better accuracy and results. The sales price is calculated for the houses using different algorithms. The sales prices have been calculated with better accuracy and precision. This would be of great help for the people.

REFERENCES

[1] 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. 169–188, 2015

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

[3] Kadir, T., & Gleeson, F. (2018). Lung cancer prediction using machine learning and advanced imagingtechniques. Translational Lung Cancer Research, 7(3), 304-312.

[4] Liu, J., Ye, Y., Shen, C., Wang, Y., &Erdélyi, R. (2018). A New Tool for CME Arrival Time Prediction using Machine Learning Algorithms:CATPUMA. The Astrophysical Journal, 855(2), 109.

[5] 99–222, Aug. 2004, ISSN: 0960-3174. DOI: 10.1023/B:STCO.0000035301.49549.88.
[Online]. Available: http://dx.doi.org/10.1023/B:STCO. 0000035301.49549.88.

[6] E.Fix and J. L. Hodges Jr, “Discriminatory analysis nonparametric discrimination: consistency properties,” DTIC Document, M. Praveena, V. Jaiganesh, International Journalof Computer Applications (0975 – 8887) Volume 169 – No.8, July 2017.

[7] Sampathkumar et al. / Procedia Computer Science 57 (2015)112 – 121.

[8] Kilpatrick, J.A Factors Influencing CBD Land Prices. Journal of Real Estate; 2000, 25: 28- 29.

[9] Wilson, I.D., Paris, S.D, Ware, J.A., & Jenkins, D.H. Residential Property Price Time Series Forecasting With Neural Networks.Journal of Knowledge-Based Systems; 2002, 15: 335-341.

[10] Mark, A.S., & John, W.B. Estimating Price Paths for Residential Real Estate. Journal of Real Estate Research; 2003: 25, 277–300.

[11] Wang, J., & Tian, P. Real Estate Price Indices Forecast by Using Wavelet Neural Network, Computer Simulation, 2005:2.

[12] David, E.D., & Paavo, M. Urban Development and Land Markets in Chennai, India. International Real Estate Review; 2008, 11: 142165.

[13] Steven, P., & Albert, B.F. Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. Journal of Real Estate Research; 2009, 31: 147-164.

[14] Sampathkumar.V and Helen Santhi.M. Artificial Neural Network Modeling of Land Price at Sowcarpet in Chennai City, International Journal of Computer Science & Emerging Technologies; 2010, 1:44–49. Available at http:// download. excelingtech.co. uk/Journal/IJCSET%20V1%284%29.pdf.