Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)
IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4

Prediction of House Pricing Using Machine Learning with Python

Mansi Jain, Himani Rajput, Neha Garg, Pronika Chawla
Department of Computer Science & Engineering
Faculty of Engineering and Technology
ManavRachna International Institute of Research and Studies, Faridabad, India - 121001
himani99raj@gmail.com, mansisahil12@gmail.com, nehagarg.fet@mriu.edu.in, pronika.fet@mriu.edu.in

Abstract. This paper 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. This paper contains what and how the house pricing model works with the help of machine learning and which dataset is used in our proposed model.

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

1. 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.

1.1 Objective

This project is proposed to predict house prices and to get better and accurate results. The stacking algorithm is applied on various regression algorithms to see which algorithm has the most accurate and precise results. This would be of great help to the people because the house pricing ids a topic that concerns a lot of citizens whether rich or middle class as one can never judge or estimate the pricing of a house on the basis of locality or facilities available. To accomplish this task, the python programming language is used. Python is a high level programming language for general purpose programming. It enables clear programming on both small and large scales. It is an easily readable language.

1.2. *Machine Learning*

Machine Learning is a field of Artificial Intelligence which enables PC frameworks to learn and improve in execution with the assistance of information. It is used to study the construction of algorithms that make predictions on data. Machine 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.

1.3. 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 Xl Rd. 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.

2. LITERATURE REVIEW

There are a couple of components that impact house costs. In this exploration, partition these components into three essential get-togethers, there are state of being, thought and territory [2]. 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 [3], 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 [4]. 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 [5].

3. METHODOLOGY

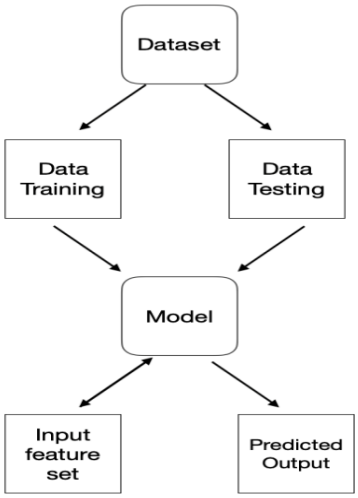


Fig.1. Block Diagram

3.1. *Data Collection*

Data collection is the process of gathering information 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.

3.2 *Data Visualization*

Data Visualization is the pictorial or graphical representation of information..It enables to grasp difficult concepts or identify new patterns. Data Visualizationis 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.

3.3. *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 express, 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 Preprocessing is a 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 information 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 preprocessing 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.

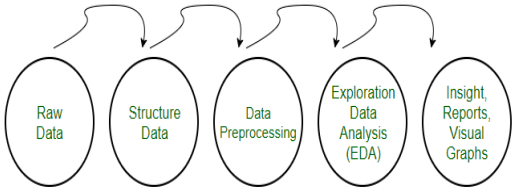


Fig.2. Data Pre-Processing

3.4. *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. The primary center ought to be on distinguishing the right qualities and discover interfaces between different information ancient rarities, for example, patterns and records.

3.4.1. *Cross Validation:*

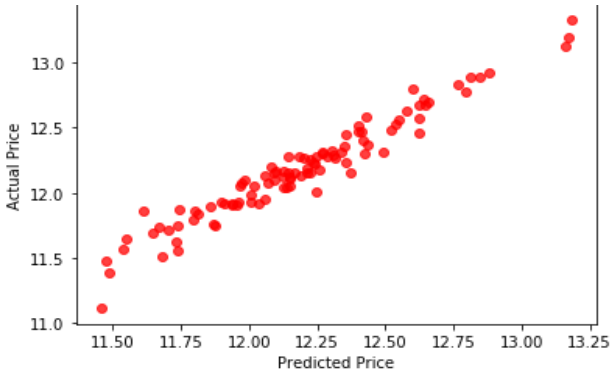
Cross validation is a strategy wherein our model is trained using the subset of the dataset and a short time later survey using the basic subset of the dataset. In validation, training is performed on 50% of the dataset and the rest 50% is used for testing purpose. The significant downside of approval strategy is that when it has been prepared for half of the dataset, it might be conceivable that the staying half may contain some helpful data which might be forgotten about at the hour of preparing the model.

3.4.2. *K-Fold Cross Validation:*

This technique splits the dataset into n number of subsets. At that point, it has been attempted for preparing on the entirety of the subsets however leave one (k-1) subset for the assessment of the prepared model. This strategy emphasizes k times with an alternate subset turned around for the preparation reason each time.

4. RESULT

To achieve the results, various data mining techniques are utilized in python language. Various factors which affects the house pricing are considered and further worked upon them. Machine learning has been considered to complete out the desired task. Firstly, data collection is performed. Then data cleaning is performed to remove all the errors from the data and make it clean. Then data pre-processing is done. Then with the help of data visualization, different plots are created, which intends to depict the distribution of data in different forms. Towards the end, the business costs of the houses were determined with exactness and accuracy. This could be achieved because a simple stacking algorithm is used to improve the accuracies of the various regression algorithms that are applied on our house pricing dataset so that they would provide better results. Apart from using the regression algorithms, some classification algorithms such as SVM algorithm, decision tree algorithm, Random Forest classifier etc. are taken into consideration and applied on our house pricing dataset. Manner which would help the people to buy houses at a reasonable cost that falls within their budgets.



5. 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.

6. FUTURE SCOPE

In future, many more algorithms can be applied on this dataset such as decision tree, Naïve Bayes, SVM etc. and find out their respective accuracies and use them to predict a better outcome and hence increase the accuracy. The KNN algorithm can also be applied to predict the accuracy. The k-means algorithm can also be applied. With the help of these algorithms, the house prices are accurately predicted. Hence, it would be of great help for the government and the people themselves. Regression algorithms are initially took up for our project but in the future, this can also be achieved using the classification algorithms. The classification algorithms can be used and it can also be applied to our house pricing dataset and see if they are being applied properly or not. The accuracy and precision of these algorithms can also be improved according to our needs. This would be of great help for the people as they would get to choose from a variety of options open up to them. They can choose the house that best suits their budgets so that they don't have to take any kind of loan from the banks. In the future, an application can also be developed for the same. That would make it even easier for the people to select the houses that best suits their budgets. Artificial intelligence can also be applied to make our project more enhanced in the future. More factors that can affect the house pricing of a particular area will also be considered. Hose pricing of an area can also depend upon political and emotional factors. Prices of houses would be more in more developed and Porsche areas where mostly wealthy people reside such as ministers and people of national interest. Buying a house can also depend upon the religious beliefs of a person. It can be affected by the direction in which the house points. People believe that facing of a house is extremely important. Some believe that the house number should make a total of either 3 or 8. All these factors are taken into consideration in the future and work upon them to make our project stronger and more relevant for public use. Although many algorithms are used in our project still, many more regression and Page classification algorithms are used to top make our project and make it more helpful for the people. Various methodologies from the field of machine learning are used to make our project more relevant. Sometimes people also prefer to stay near areas where basic facilities are easily available such as a general store, mother dairy, photocopy shop etc. This is also an extremely important factor that may affect the prices of houses and can be taken into consideration in the future. All the major factors that can affect the prices of houses in a particular area are almost covered and have worked upon them. In the future, some of the minor factors that can affect house pricing on a smaller scale can be identified and can work upon them that how do they affect house pricing and what can be done to minimize it. In the future the model deployment of more algorithms can be performed to achieve accurate results.

7. REFERENCES

[1] Jain, N., Kalra, P., &Mehrotra, D. (2019). Analysis of Factors Affecting Infant Mortality Rate Using Decision Tree in R Language. In Soft Computing: Theories and Applications (pp. 639-646). Springer, Singapore.

[2] [2] 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

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

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

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

[6] Velankar, S., Valecha, S., &Maji, S. (2018, February). Bitcoin price prediction using machine learning.In Advanced Communication

Technology (ICACT), 2018 20th International Conference on (pp. 144-147).IEEE.
[7] Malhotra, R., & Sharma, A. (2018). Analyzing Machine Learning Techniques for Fault Prediction Using Web Applications.Journal of Information Processing Systems, 14(3).
[8] Choo, M. S., Uhmn, S., Kim, J. K., Han, J. H., Kim, D. H., Kim, J., & Lee, S. H. (2018). A Prediction Model Using Machine Learning Algorithm for Assessing Stone-Free Status after Single Session Shock Wave Lithotripsy to Treat Ureteral Stones. The Journal of urology.
[9] Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L., & Farahmand, M. (2018). A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. Biocybernetics and Biomedical Engineering, 38(1), 1-15.
[10] Fan, C., Cui, Z., &Zhong, X. (2018, February). House Prices Prediction with Machine Learning Algorithms.In Proceedings of the 2018 10th International Conference on Machine Learning and Computing (pp. 6-10).ACM.
[11] Zhou, J., Zhang, H., Gu, Y., &Pantelous, A. A. (2018). Affordable levels of house prices using fuzzy linear regression analysis: the case of Shanghai. Soft Computing, 1-12.
[12] Jang, H., Ahn, K., Kim, D., & Song, Y. (2018, June). Detection and Prediction of House Price Bubbles: Evidence from a New City. In International Conference on Computational Science(pp. 782-795). Springer, Cham.
[13] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms.Pattern recognition, 30(7), 1145-1159.
[14] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data.Expert Systems with Applications, 42(6), 2928-2934.
[15] Harrison, D., and D. L. Rubinfeld. 1978. “Hedonic Housing Prices and the Demand for Clean Air.”J. Environ. Econ. Manag.5 (1): 81–102.