ISSN: 2581-3404 (Online) IF: 5.68 (SJIF)
IJIRTM, Volume-8, Issue-6, November-2024.

House Price Prediction Using Machine Learning and Ensemble Techniques

Shagun Tiwari[1], Priyanka Bhatele[2]
[1]Sagar Institute of Science Technology and Research, Bhopal, India, 462044
[2]Sagar Institute of Science Technology and Research, Bhopal, India, 462044
[1]Shagun.tiwarii1999@gmail.com, [2]priyankabhatele@sistec.ac.in

Abstract. *Today all business is totally depending upon Data. At Every second Petabytes of data are generation from different domains or areas. We all Knows that utilization of data is useful data is all around. Today's scenario many Top-Notch Companies are depends upon data. They are growing day by day in their business area based upon the net profit. Data mining is really a important tools for processing the data from its Raw format to useful data from last many years. Recently data mining is migrated to Machine Learning. Machine Learning is a branch or sub domain of Artificial Intelligence and Mathematical Statistics. Machine learning can be divided into supervised, unsupervised and reinforcement learning. Here Researcher's wants to find such types of solution which can find for the solution for People who want to buy a new house are more conservative with their budget and market strategies. In the many available current system, the calculation of house price is calculating without the necessary forecasting about future market business trends. The goal of this Paper is to predict efficient house pricing for real estate customers in terms of their budgets and requirements the trends and price limits of the previous market, and near coming future developments. The outcome of this Paper considers client specifications and then combines the application of several Machine Learning Algorithms like regression algorithms for finding our Analysis. This will help customers to invest in an asset without an agent and it also reduces the risk involved in the transaction.*

Keywords: - Classification Algorithm, Machine Learning, Data mining, House price forecasting, Prediction, Linear regression, Random Forest, Gradient boosting.

Introduction

In Recent time House prices are quickly changed in many area we can say that this change goes like dramatically, Due to this changes modern many economic zone are created and varies their aspect . If we are really good in accurate prediction model then that will be very helpful in the interest of all real estate market applicants such as property sellers, property buyers, investors, and many more participants who directly and indirectly involve in this Real Estate Business. We took a case study of a city Melbourne that has been recognized as the world's most suitable city for staying according to the many organizations it has been found that many new migrants from all over the world is coming to this city and wants to buy the properties to live here. As a result of interest of many, the price in this city is continuously increasing

over the past few years and investing in any properties is generally secure and beneficial for all who invested here [1]. In recent days, thanks to the huge evolution in data generation & accumulation, artificial intelligence (AI) driven decision making is really very good in many aspects. Machine learning related algorithms is using for in depth data analysis [2] and their useful applications have been broadly used in both industry and education domain for further technical and domain research [3], [4]. We all knows that Regression is very strong mathematical model which has two major sub part i.e., Linear regression and logistic regression which are typically classified as supervised learning techniques [7], [8]. In the supervised learning process, a dataset that contains the desired outputs is first used for model training and the well-tested system that is capable of processing new given inputs to generate or predict very precise outputs. LR is generally used for solving regression problems whereas logistic regression is utilized in data classification [9]. Even though Predicting the housing price data has been very often used in linear prediction algorithms, partially the linear regression method, in many online data science forums or lectures.

In this paper, the useful detailing of both the LR (Linear Regression) and the logistic regression are explained thoroughly. If we talk about, absolute datasets from the city i.e., Melbourne used to find the performance of the trained models. In this finding, the housing price prediction dataset of Year 2017 is used for training of model and the dataset of Year 2018 is used for testing the model of Price prediction mechanism. Here 1st research paper, the linear regression algorithm is used to model the relationship between the price of the house and its different features or property like distance to the city of Melbourne. It defines a simple linear model can absolutely predict the average sold price of the properties if the training dataset contains huge amount of data. In the 2nd section of this paper, the logistic regression model is utilized for classify the sold houses which belong to different councils based on both the price and the distance information. It shows that when the considered two councils are far away from each other, the classification accuracy is 100%. When the considered two councils are close to each other, the performance measures i.e., accuracy is still above 85%.

The remaining of this paper is organized as follows. The dataset which is utilized for the analysis and its prior-processing are first discussed in 2nd section. 3rd Section and 4th Section is used to describe the information of linear regression and logistic regression models in detail.

Predictive models- Here the relationships and some previous patterns that usually edge to a certain behaviour. By explaining the descriptive variables, we can judge the prediction results in the dependent variables.

Descriptive models- Here the partition into smaller part that is known as segment. Usually, it is used to classify.

1.1 Machine Learning Techniques in House Prediction Mechanism

Now days Many Industries is opting Machine Learning Mechanism for their In-depth Analysis. We all knows that India is developing Countries Here Real Estate Industries is very important aspect for many cities, here 1st step for development is Infrastructure Development Where we have Roads, Houses, Parks, Hospitals and market Complexes. We are observing that many SEZ Zone is developing for evolving the city. If people start living their many industries is growing automatically.

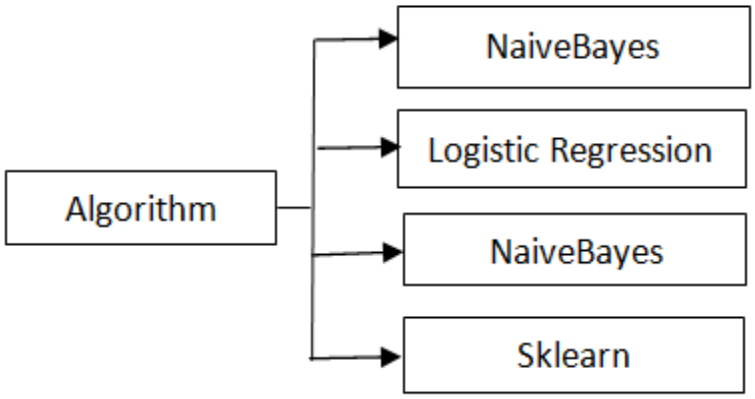

Figure 1: ML Algorithms

1.2 Overview of House Prediction System

House Prediction System is used to analysis the cost prediction on the basis of many parameters some so them is listing below:

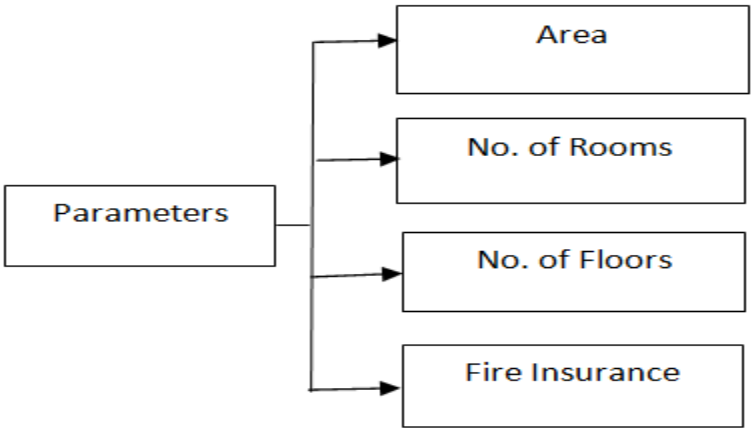

Figure 2: Parameters

In other ways we can say that Price prediction depends upon many features.

II. Literature Review

Here Author's explained in their paper uses of the greedy price evaluation model. The modelling is implemented for the city of Georgia, for finding the house valuation in recent days and how much it will be increased in upcoming years. If we go in-depth in this paper, we found that a very good linking between the price variation of any available properties in terms of number of available rooms in our Houses the garage area, area of bathrooms, how our houses is prevented from fire, and overall living area based on the square meters [6]. The target of this analysis is on the use of machine learning algorithms in the real estate field. With regression analysis and particle swarm optimization, the objective of the study is to predict house prices totally dependent on housing city. Here Researcher's used swarm optimization

mechanism for calculating the effective variables, and find out the optimum solution using regression Mechanism. The findings of this study have been shown to be acceptable by adding two algorithms i.e regression and PSO [6]. Here Researcher's used; support vector Regressor is implemented to estimate China's housing prices from 1993 to 2002. In the support vector Regressor model is used to tune the hyperparameters. The error scores achieved for the genetic algorithm was usually lesser than 5%. The proposed model measures image features using PCA i.e., part of Artificial Intelligence. The prediction measures i.e. Precision, MSE between the expected results and the given dataset. This dataset is provided by Japan real estate service provider [8].

The dataset which is used here for analysis and its pre-processing are first discussed in 1st Part. Part 3rd and 4th explained the studied linear regression and logistic regression models in detail. In part 5th & 6th both the training Mechanism & the testing mechanism are presented. Finally, 7th part concludes this paper. Apartment's values totally depend upon the location where that apartment situated [10]. Researcher's done many Reviews for finding of multi-family housing plane in different cities as research [11].

In this paper Researcher's focuses on 4 different regression techniques and an ensemble mechanism by adding or combining K-Nearest neighbor and Random Forest Technique to find out the predicted cost of any apartments. The ensemble mechanism can estimate the prices with a minimum error of 0.0985 and remains constant when researchers implemented a technique like PCA. The 4 different regression techniques used are LR (Linear Regression), SVM (Support Vector Machine), RFR (Random Forest Regression, & k-nearest neighbor [13].

III. Objective

The motivation of House Prediction through machine learning is due to the need of market and huge data available for this domain. We know that Availability of huge data force us to shift from conventional programming to Machine Learning concept. Here frequent change of data and requirement motivate to select Ensemble Learning Algorithms. Here we are trying to analysis the performance of some of the existing classification algorithms like Gradient Boosting, Random Forest and Linear Regression. For claiming their performance, we will find the Accuracy and many more parameters. We are taking Kaggle Portal for the analysis of our Dissertation Work.

IV. Problem Identification

Classification is the finding rules that partition the data into different groups. Apply classification in training dataset. Analyse the training dataset and develop a model based on the class label and assign a class label to the future unlabelled records. Classification technique in this class field is known as supervised learning. This technique is used to develop a better classification process, it can be used to classify data and for the better understanding of each class of the dataset. There are many classification algorithms which will be used Naive Bayes algorithm, support Vector machine, K Nearest Neighbour, logistic regression, decision trees, neural network. Also, will work on linear and geometric analysis. These mechanisms are used in credit cards, fraud detection, banking, medical etc. Some classifications which are used in this dissertation are: Classification: Train, Validation, and Test Split.

V. Flow Diagram

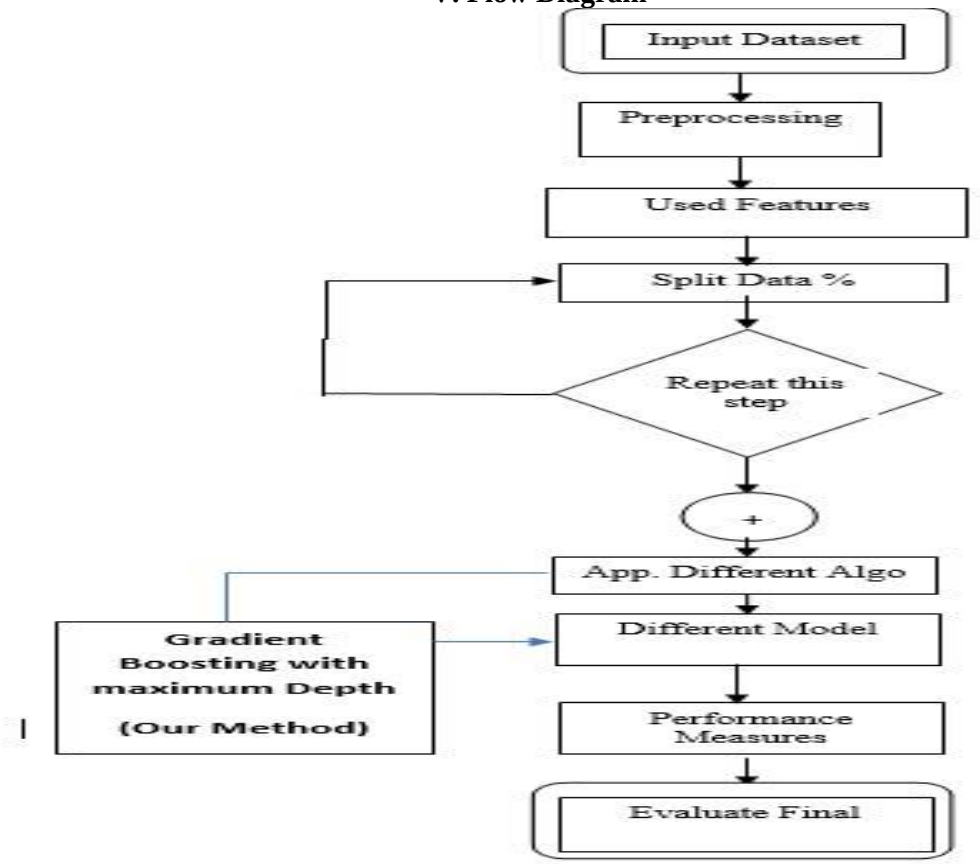

Figure 3: Flow Diagram

VI. Algorithms

Step 01: Store Data from Repository (Kaggle or UCI Machine Learning)
Step 02: Import Prior Libraries
Step03: Apply pre-process for better understanding for our Data Set
Step04: Divide our Data set into two parts names as
Step05: Apply Feature Extraction
Step 06: Visualize Data for better understanding with the help of matplotlib
Step07: Repeat step04 & step05 Multiple Times for Better Understanding of data
Step08: Here we will apply machine learning algorithms for better prediction in this work we applied different algorithms:
Step09: Apply Different Model Individually and calculate the performance value
Step10: Finally Compare Results on performance parameters like Accuracy and claimed that Gradient Boosting Gives Better Performance.

VII. Experimental Results & Discussion

The given table show that how different algorithms give different result at different data splitting process.

Table 1: Output Results

S. No	Training	Testing	LR	RFR	GBR
1	60	40	88.57	88.76	90.22
2	65	35	88.14	88.06	89.63
3	70	30	87.56	88.07	89.15
4	75	25	88.72	88.66	90.76
5	80	20	89.26	89.39	91.81

In Table 1 Researcher's explained How Different Algorithms gives different performance when we split our data during model creation. In Table 6.1 we found that firstly values of Accuracy of different algorithms increases but after some split it goes down and again values are increasing when split of data changes. Finally, we found that maximum performance of accuracy given by different algorithms when split percentage is 80-20. We know that in machine learning model creation it plays very important roles for model creation. In Table 1 they used following short forms which is given below:

LR=Linear Regression
RFR= Random Forest Regression
GBR=Gradient Boosting Regression

Figure 4: Graph between Algorithms

In Figure 4 Researcher's explained how different algorithms gives different result when we splitting the Data. In figure 6.1 At first position we split data into 60-40 , At second position we split data into 65-35, At third position we split data into 70-30 , At fourth position we split into 75-25 , At fith position we split data into 80-20.

sklearn.model_selection.train_test_split(*arrays, **options)
test_size : float or int, default=None
train_size : float or int, default=None
random_state : int or RandomState instance, default=None
shuffle : bool, default=True

Note: In above formula which we used in our implementation during data spliting. Random state parameter can do much effect.

VIII. Conclusion And Future Scope

We conclude the observation about the previous technique. Our observation different terms and condition. This represents our work in a new approach. In our work we have tried to improve Accuracy we applied ensemble techniques where we applied two different algorithms i.e., Random Forest algorithm and Gradient Boosting Regressor algorithm in every algorithm we got different Accuracy goes to 90 %. Again, we adjust training and testing splitting by this mechanism we get better result. For finding those facts we are doing following in our System:

- Firstly, Fetched the recent Data through Kaggle Repository.
- Fetched Data in CSV format where we have numbers of attributes to better understanding.
- For preprocessing Fetched data convert it into Excel Format for better processing
- Apply multiple algorithms and record their Accuracy
- Finally fetched Accuracy of different algorithms & represent with their numerical values and convert it into graphical format.

The future works focus on applying some other techniques to improving the performances of these methods for up to maximum extent. After research and implemented our work, we find that many works can be enhanced their values and performance of given research topics out of them some points are given below:

- Neural network may enhance their result because every stage or HOP NN improve their result.
- In near future we will apply Deep Learning concept which may enhance their accuracy.
- Finally, in future we will try to apply comparisons upon different Dataset with same Algorithm.
- In future Authors can try for different preprocessing library to see their effect.

References

[1] Chenchen Fan , Zechen Cui , Xiaofeng Zhong, “House Prices Prediction with Machine Learning Algorithms,” , ICMLC 2018, February 26–28, 2018, Macau, China © 2018 Association for Computing Machinery. ACM ISBN 978-1-4503-6353-2/18/02…$15.00 DOI: http://dx.doi.org/10.1145/3195106.3195133.

[2] Ayush Varma , Sagar Doshi , Abhijit Sarma , Rohini Nair, “House Price Prediction Using Machine Learning And Neural Networks, “Computer Engineering Department Somaiya College Of Engineering,Vidyavihar, Mumbai- 400077

[3] Li Li and Kai-Hsuan Chu, “Prediction of Real Estate Price Variation Based on Economic Parameters,” Proceedings of the 2017 IEEE International Conference on Applied System Innovation IEEE-ICASI 2017 - Meen, Prior & Lam (Eds).

[4] Muhammad Fahmi Mukhlishin, RagilSaputra, Adi Wibowo, "Predicting House Sale Price Using Fuzzy Logic,Artificial Neural Network and K-Nearest Neighbor," 2017 1st International Conference on Informatics and Computational Sciences (ICICoS) 978-1-5386-0903-3/17/ © 2017 IEEE .
[5] The Danh Phan, "Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia,"2018 International Conference on Machine Learning and Data Engineering (iCMLDE) 978-1-7281-0404-1/19/$31.00 ©2019 IEEE DOI 10.1109/iCMLDE.2018.00017 .
[6] Shantanu Chakraborty, Tomonobu Senjyu, "Application of Incentive Based Scoring Rule Deciding Pricing for Smart Houses," 978-1-4799-1303-9/13/$31.00 ©2013 IEEE.
[7] Parasich Andrey Viktorovich, Parasich Viktor Aleksandrovich, Kaftannikov Igor Leopoldovich, "Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning," The work was supported by Act 211 Government of the Russian Federation, contract № 02.A03.21.0011
[8] Nehal N Ghosalkar, Sudhir N Dhage, "Real Estate Value Prediction Using Linear Regression," 978-1-5386-5257-2/18/$31.00 ©2018 IEEE
[9] Dr. Swapna Borde, Aniket Rane, Gautam Shende and Sampath Shetty,"Real Estate Investment Advising Using Machine Learning," IRJET, 2017.
[10] "Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile," AMSE Conference Santiago, Chile, 2016.
[11] Mansurul Bhuiyan and Mohammad Al Hasan, "Waiting to be sold: Prediction of Time-Dependent house selling probability," IEEE International Conference on Data Science and Advanced Analytics, 2016.
[12] Wan Teng Lim, LipoWang,Yaoli Wang and Quing Chang, "Housing Price Prediction Using Neural Networks," IEEE 12th International Conference on Natural Computations, Fuzzy Systems and Knowledge Discovery, 2016.
[13] Muhammad A. Razi and Kuriakose Athappilly, "A comparative predictive analysis of neural networks (NNs), nonlinear regression and classication and regression tree(CART) models," Western Michigan University,2005.
[14] Youness El Hamzaoui and Jose Alfredo Hernandez Perez, "Application of articial neural networks to predict the selling price in the real estate valuation process," Morelos, Mexico, 10th Mexican International Conference on Articial Intelligence, 2011.
[15] Ruben D. Jaen, "Data Mining: An empirical Application in Real Estate Valuation," Florida International University, 2002.
[16] D. P. W. Ellis and G. E. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," Proc. Int. Conf Acoustic, Speech and Signal Processing, Honolulu, HI, 2007.