Research Article

House Price Forecasting Using Machine Learning Methods

R.Monika[1], J.Nithyasree[2], V.Valarmathi[3], Mrs.G.R.Hemalakshmi[4*], Dr.N.B.Prakash[5]

[1,2,3]UG Students, Department of CSE, National Engineering College, K.R.Nagar, Kovilpatti-628503
[4]Assistant Professor (Sr.Grade), Department of CSE National Engineering College, K.R.Nagar, Kovilpatti-628503
[5]Associate Professor, Department of EEE ,National Engineering College, K.R.Nagar, Kovilpatti-628503

Article History: Received: 11 January 2021; Revised: 27 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract - Machine learning has a huge impact in the previous years in picture recognition, spam redesign, typical discourse order, and item proposal. The current Machine learning calculation causes us to redesign security cautions, guaranteeing public wellbeing, and improving clinical improvement. By analyzing many algorithms, we decide to engage in machine learning. Determining the price of the house is vital these days as the price of the land and the price of the house increases each year. This utility can assist clients to make investments in a property except drawing near an agent. Here we use various regression methods of supervised machine learning using Python. Regression is used to predict future values based on the independent variable. The model's evaluation is done by calculating the error value. When a small error occurred, it would give great accuracy to our regression model, so in this problem, we are going to predict the house values. The following algorithms Extreme Gradient Boosting, Gradient Boosting Regression, Random forest regression, Light Gradient Boosting Machine regression, support vector regression were used to forecast the house values. Further, these algorithms are compared according to the predicted results. The eventual outcome will be shown as the best calculation as far as forecast exactness.

Keywords: Machine learning, house price, prediction, regression

I. Introduction

This article alludes along with the most recent forecast on research expectations considering patterns to additional arrangement of their financial matters. The principle inspiration of our project is to predict the house price variations forecasting utilizing the machine learning algorithms and discovering the reasonable anticipating price with a low error rate. This is an intriguing issue in light of the fact that a great people will in the long run purchase. This problem makes us to study real estate advertise and helps with knowledgeable decisions. The examination that was done in this work is primarily found on the datasets of the USA.

For the buyers of land properties, and automated value forecast structure can be helpful to discover under/over rated properties presently available. The information will be very much valuable for the initial buyers with low or zero experience and propose buying techniques for purchasing properties.

Machine learning based algorithms was implemented in orders like mechanical designing, bioinformatics, clinical, drugs, actual science, and measurements to gather data and figure prospect occasions. In the present developmental growth of housing market, machine learning plays a vital role to forecast the property price.

There are diverse machine learning algorithms to forecast house prices. In our project work, attempt has been made to deliver probable regression techniques appropriate to the issue. This venture will use the following regression techniques XGboost, Gradient Boosting Regression, Random forest regression, lightGBM regression, support vector regression.

II. Literature survey and review statement

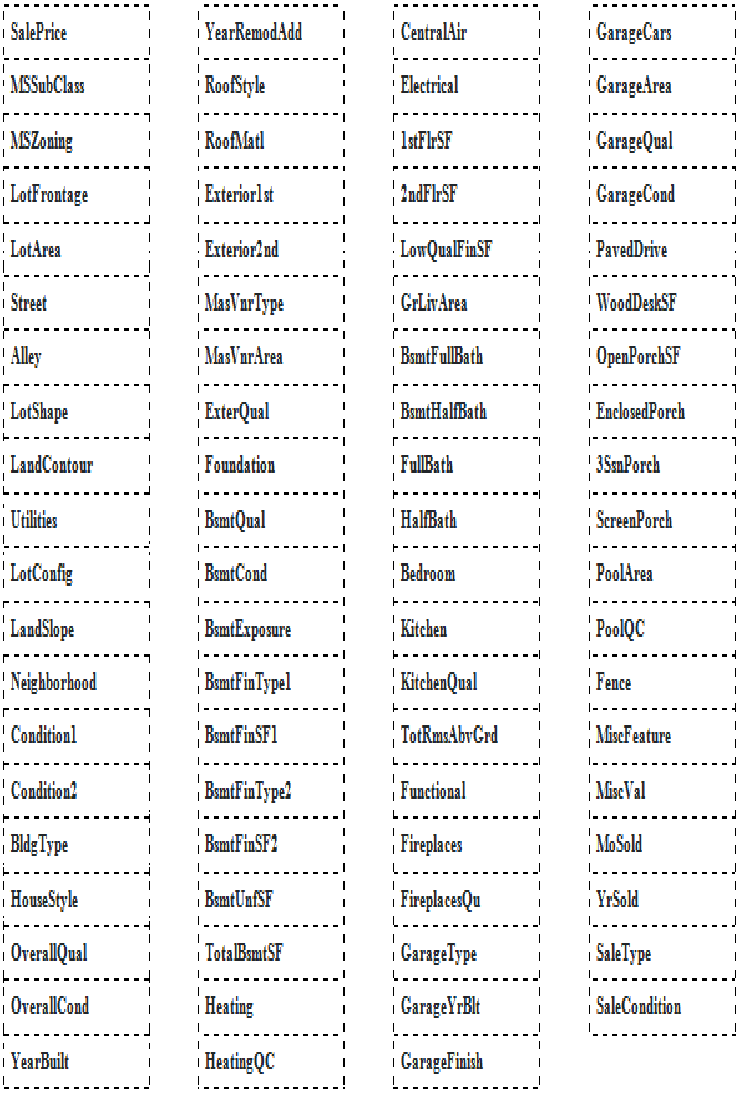
Byeonghwa Park[1], implemented machine learning algorithms for the housing price prediction accuracy. The housing data was analyzed from townhouses in Fairfax Country and compares the classification accuracy performance of various algorithms. To help a real estate agent, he then develops a better prediction model for enhanced decision based on house price assessment. **CH.Raga Madhuri, Anuradha G[2],** estimated house price by the analysis of fare ranges, foregoing merchandise and forewarns of developments. The author discussed diverse regression techniques such as Gradient boosting and AdaBoost Regression, Ridge, Elastic Net, Multiple linear, LASSO to locate the most excellent. The performance measures used are [MSE] Mean Square Error and [RMSE] Root Mean Square Error. In the comparative study, the gradient boosting regression technique is best fitted. **Feng Wang, Ynag Zoe, Haoyu Zhang and Haodang Shi [3],** the tensor flow framework is implemented to predict the residence price based on deep learning. The experimental dataset is accessed from the internet by using Scrapy. Relative investigations were directed among the tensor flow approach and the SVR method. The results shows superior compared to SVR method.

Zhongyun, Jiang, Guoxin, Shen[4], develops 6-layer BP neural network with Kera's deep learning. For backend Tens flow or Theano is used. The test results gives the real value of home price as 95.59%. The Gaussian noise model is favorable for the house price forecast. **Aboozar Taherkhani , Georgina Cosma[5],** In this paper, the capacity of Adaptive Boosting (AdaBoost) is coordinated with a Convolutional Neural Network (CNN) to plan another machine learning technique, AdaBoost-CNN, which can manage enormous imbalanced datasets with high exactness. The proposed AdaBoost-CNN is intended to decrease the computational expense of the old-style AdaBoost when managing enormous arrangements of preparing information by diminishing the necessary number of learning ages for its fixing assessor. The exhibition of AdaBoost-CNN on multi-modular information examination will be explored in future exploration where various modalities of information will be prepared by various CNNs. **Zhen Peng, Qiang Huang, Yincheng Han[6],** precisely study the cost of recycled houses, and examined 35417 bits of information caught by the Chengdu HOME LINK organization. First and foremost, the caught information was cleaned, and the attributes were chosen. The test results show that the precision of XGboost expectation is the most elevated, and the forecast exactness score arrives at 0.9251. Dissimilarity with linear regression, decision tree model, the XGboost algorithm gives improved speculation capacity and strength in information forecast, and, furthermore, data prediction over-fitting aspect, establishing a strong framework for the ensuing recycled house value expectation.

Cun Jia, Xiunan Zoua, Yupeng Hub, Shijun Liub[7], developed Shapelet-based strategies pulled in a ton of consideration in the most recent decade. As the time intricacy of the shapelet determination is excessively high, a great deal of quickening techniques is implemented i.e an XGBoost classifier dependent on shapelet highlights is utilized in XG-SF to improve order exactness. **Qiansheng Zhang, Ziqi Wu[8],** Contrasted with a single Decision Tree, boosting algorithms work better at handling lost information as well as not all that precise to noisy information. Then again, boosting algorithms work without weaknesses of a single Decision Tree which creates a complex non-direct relationship by utilizing a distinctive return tree as a weak returning gadget, making the precision higher. It can evade over-fitting by controlling emphasis with the goal that the speed of computation can be improved. Simultaneously, with a bagging algorithm and a random forest algorithm, a boosting algorithm makes powerless returning gadgets more adjusted. Furthermore, the stock forecast cost can be altered by utilizing the fluffy chance assumption hypothesis. At long last, the exact examination of stock value expectation for PING A BANK of China by using the upgraded boosting relapse calculation shows that MSE and MAPE forecast deviations decline incredibly, which improves the expectation of execution of boosting calculation without ideal contentions.

III. Dataset Collection and Preprocessing

In this project, the dataset taken are from a kaggle application. It is an open source platform. This dataset involves 79 explanatory variables representing every aspect of homes in King Country, United States of America.



SalePrice
MSSubClass
MSZoning
LotFrontage
LotArea
Street
Alley
LotShape
LandContour
Utilities
LotConfig
LandSlope
Neighborhood
Condition1
Condition2
BldgType
HouseStyle
OverallQual
OverallCond
YearBuilt

YearRemodAdd
RoofStyle
RoofMatl
Exterior1st
Exterior2nd
MasVnrType
MasVnrArea
ExterQual
Foundation
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinSF1
BsmtFinType2
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
Heating
HeatingQC

CentralAir
Electrical
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
Bedroom
Kitchen
KitchenQual
TotRmsAbvGrd
Functional
Fireplaces
FireplacesQu
GarageType
GarageYrBlt
GarageFinish

GarageCars
GarageArea
GarageQual
GarageCond
PavedDrive
WoodDeskSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
PoolQC
Fence
MiscFeature
MiscVal
MoSold
YrSold
SaleType
SaleCondition

Fig. 1 Dataset Variables

From the above figure.1, we can notice the independent variables from the housing dataset are the explanatory variables. The dependent variable is SalePrice remaining variables are independent variables. We have 1460 perceptions of 80 features in the training data frame. In the following process, we will find the top 10 correlation variables matrix. The below diagram figure.2 shows the top 10 correlated variables.

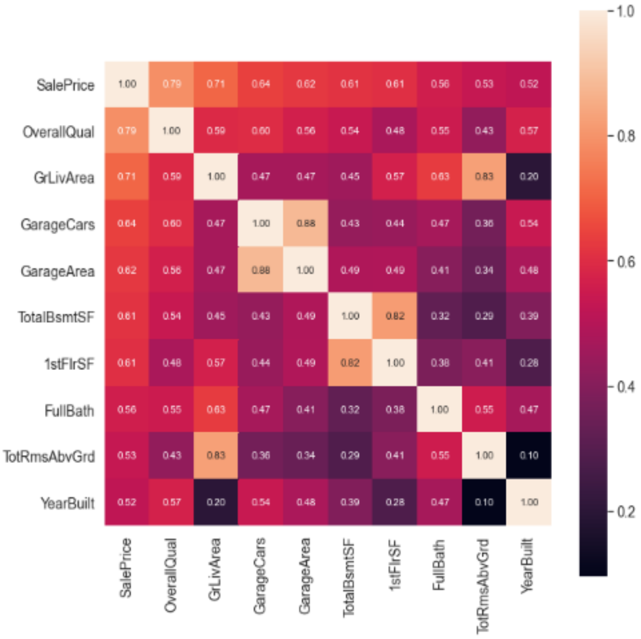
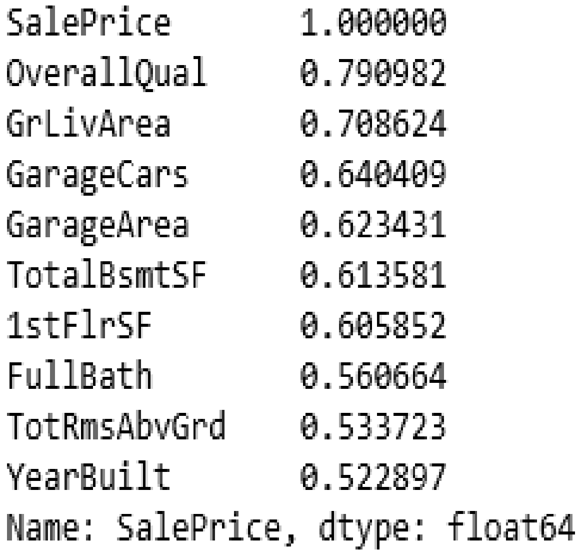


Fig. 2 Correlation Matrix

The Value of correlation for the top 10 correlated variables is as given in figure.3



```
SalePrice       1.000000
OverallQual     0.790982
GrLivArea       0.708624
GarageCars      0.640409
GarageArea      0.623431
TotalBsmtSF     0.613581
1stFlrSF        0.605852
FullBath        0.560664
TotRmsAbvGrd    0.533723
YearBuilt       0.522897
Name: SalePrice, dtype: float64
```

Fig. 3 Correlated Variables

In the next step, handling the missing values in dataset and then fill missing values in dataset where in the numerical columns fill with 0 and categorical columns with ‘None’. The below Table.1 displays the packages needed to implement this project.

Table I
Packages and versions

Requirements	Versions
Pandas	1.1.5
Numpy	1.19.5
Matplotlib	3.2.2
Seaborn	0.11.1

Xgboost	0.90
Scikit-learn	0.24.1

IV. Aim and objective of the study

The fundamental target of the undertaking is to figure house costs utilizing Machine learning algorithms. Using five different models we are predicting the house price. And finally, the best fitting algorithm is shown based on the prediction accuracy terms.

(i) Extreme Gradient Boosting:

The short form of Extreme Gradient Boosting is called XGboost algorithm. It can be viewed as a variation of the GBDT calculation [6]. XGboost expanded and enhanced GBDT has a more rapidly algorithm and superior precision.

(ii) Gradient Boosting Regression

A Gradient Boosting Machine or GBM merges the predictions from multiple decision trees to generate the final forecast. This ML technique is used for regression and classification problems.

Working of GBM: GB includes three parts; Loss work, Weak Learner, Additive Model.

Loss function: This function depends on the problems being solved. Numerous standard capacities are upheld, and you can characterize your own. The upside of the incline boosting framework is that another boosting estimation shouldn't be resolved for each incident work that may be used. All the things are equal; it is a nonexclusive enough structure that differentiable misfortune capacity can be utilized.

Weak Learner: In Gradient boosting algorithm shown in figure.4 Decision trees are utilized as the feeble student. By picking the best part focused on virtue the trees are developed in an eager way. This is to build up that the students stay weak, however can in any case be inherently a covetous way.

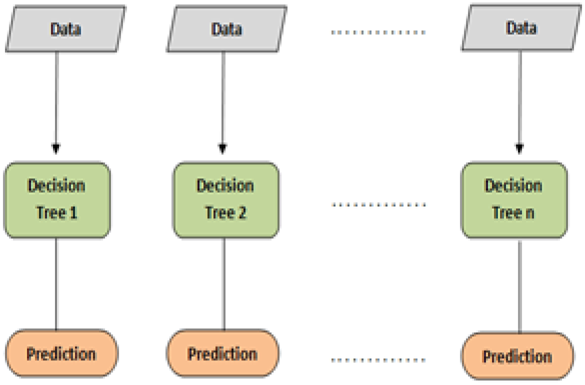


Fig. 4 Gradient Boosting Algorithm

Additive Model: The current trees in the model continue as before when new trees are added each in turn. While adding the trees, an angle drop methodology is utilized to diminish the misfortune. Generally, gradient descent is utilized to limit a bunch of boundaries, like the coefficients in a relapse condition or loads in a neural network. Ensuring to figure out misfortune or mistake, the heaps are revived to restrict that blunder.

(iii) Random forest regression

Random forest regression shown in figure.5 is a sort of supervised machine learning dependent on ensemble learning. Ensemble learning is a kind of realization where you join various sorts of calculations or a similar calculation on numerous occasions to shape an all the more remarkable forecast model. The arbitrary timberland calculation consolidates various calculations of a similar kind for example numerous choice trees, bringing about a woods of trees, subsequently the name "Arbitrary Forest". The irregular timberland calculation can be utilized for both relapse and characterization errands.

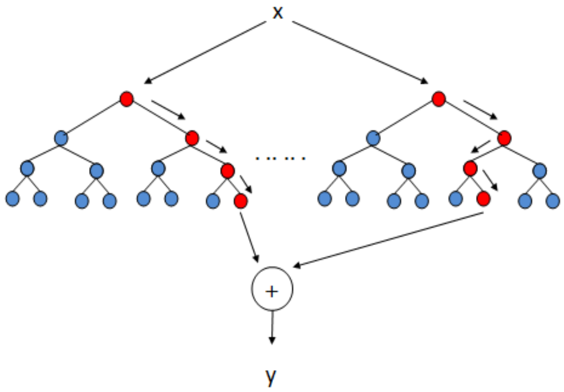


Fig. 5 Random Forest Regression

(iv) Light Gradient Boosting Machine regression

LightGBM algorithm shown in figure.6 incorporates a few boundaries, named hyper parameters. The hyper parameters altogether affect the presentation of LightGBM algorithm. They are ordinarily set physically and afterward tuned in a consistent experimentation measure. In our work, a hyperparameter optimization calculation is smartly incorporated.

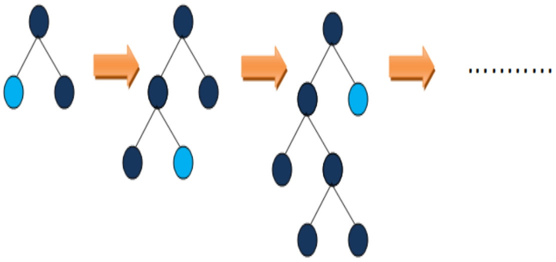


Fig. 6 LightGBM Regression

(v) Support Vector regression

Support Vector Regression (SVR) illustrated in figure.7 utilizes a similar standard as SVM, however for regression issues. How about we put in no time flat understanding the thought behind SVR. The issue of regression is to discover a capacity that approximates planning from an info area to genuine numbers based on a preparation test[21][22][23]. So we should now plunge profoundly and see how SVR functions really**.**

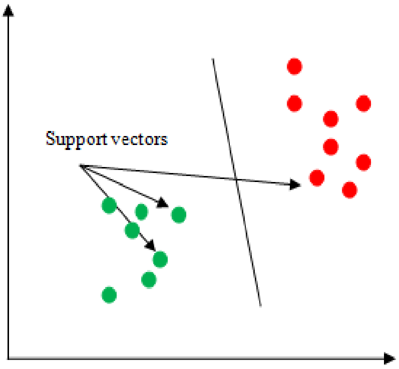


Fig. 7 Support Vector Regression

V. Implementation works

This exploration utilizes jupyter IDE. It conveys gear for measurement cleaning, change of information, a reenactment of numeric values, modeling the utilization of insights, perception of measurements, and PC becoming more acquainted with apparatuses. In this work, collection of house deals-related records to gauge the home charges fundamentally dependent on the genuine world dataset ruler area is done. All the above referred to relapse techniques are completed utilizing the above unmistakable devices. To find the climate agreeable relapse approach for expectation, we require sure boundaries to examine the strategies. The performance measures taken are [MSE] Mean Square Error and [RMSE] Root Mean Square Error. Table.2 addresses the abstract of the boundaries during the implementation of the algorithm practically.

Table II
Comparison of Algorithms

Algorithm	Score	MSE	RMSE
Extreme Gradient Boosting	0.898990	1.192018e+10	109179.574213
Gradient Boosting Regression	0.893702	1.254429e+10	112001.312071
Random Forest Regression	0.879932	1.416928e+10	119034.779432
Light Gradient Boosting Machine Regression	0.906766	1.100254e+10	104893.008553
Support Vector Regression	0.900679	1.172093e+10	108263.256764

From the above table, we can effortlessly operate comparison of exclusive algorithms without a doubt to locate the first-rate amongst them. Figure.8 to Figure.12 beneath is utilized to unmistakably picture the presentation of different procedures. In all the figures, x-pivot addresses the different relapse strategies considered for study and y-hub addresses the score esteems noticed.

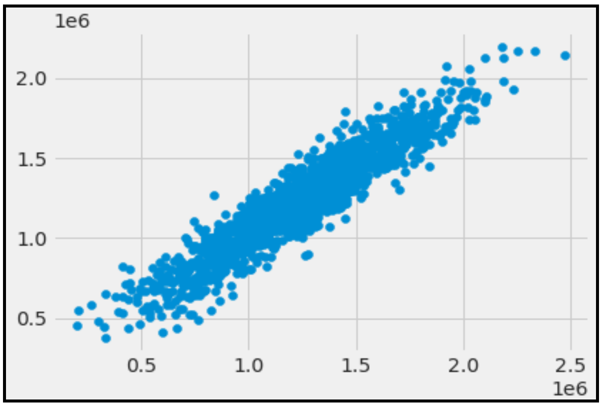


Fig. 8 Random Forest Regression

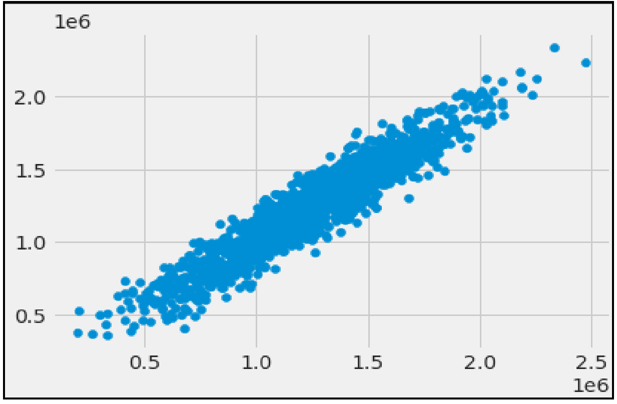


Fig. 9 Gradient Boosting Regression

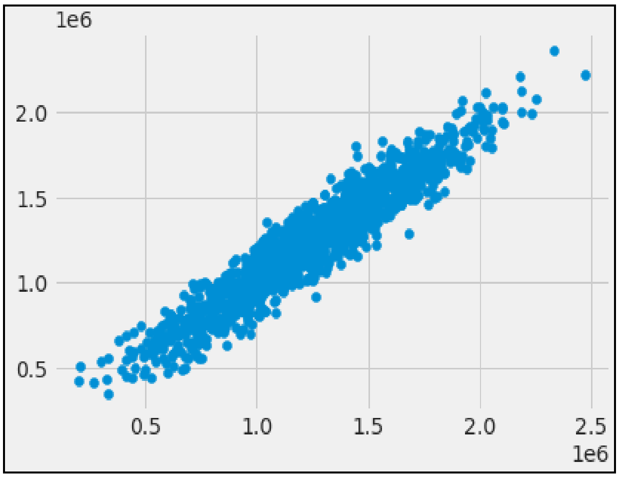


Fig. 10 XGBoost

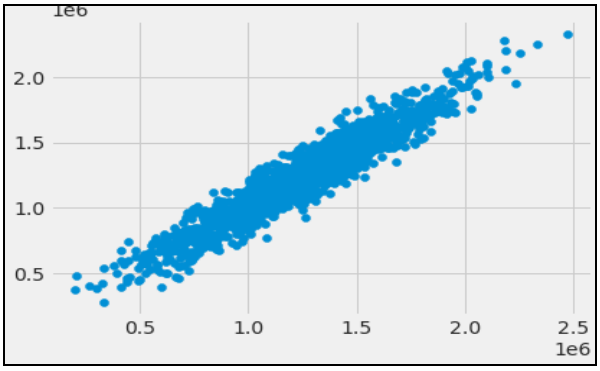


Fig.11 LightGBM Regressor

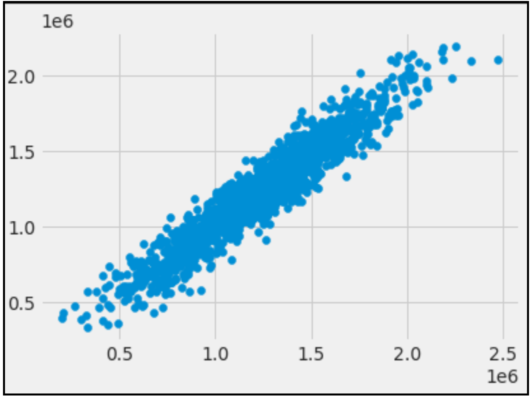


Fig. 12 SV Regression

The graphical portrayal of all the distinctive relapse procedures recorded above is unmistakably addressed underneath using Matplot material.

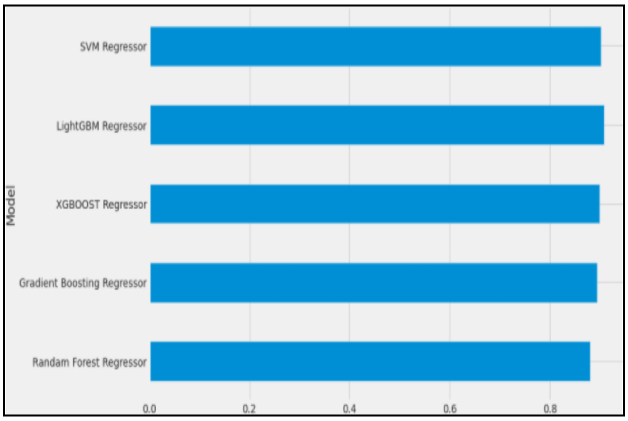


Fig. 13 Comparison Graph

The above figure.13 displays the comparison among the five models. The Graph is plotted by using score of all the regressions. Among them LightGBM Regressor has high accuracy.

VI. Results and Discussion

The whole work is mainly concentrated on the analyzing the various machine learning algorithms (Extreme Gradient Boosting, Gradient Boosting Regression, Random forest regression, Light Gradient Boosting Machine regression, support vector regression) with respect to house pricing with King County publicly available Dataset. From the examination study, the LGB machine regression has high accuracy using the performance measures [MSE] and [RMSE].

VII. Conflict of Interest

There is no conflict of interest in publishing this paper

VIII. Acknowledgement

The authors thank the National Engineering College Management, Director, Principal, HOD/CSE and HOD/EEE for having granted permission to carry out the project work at the institution utilizing the infrastructure and software facilities. The authors also thank the CSE department project coordinator and faculty members for having reviewed and provided valuable suggestions towards successful completion of the work.

IX. References

1. Byeonghwa Park , Jae Kwon Bae “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data”. 2017
2. CH.Raga Madhuri, Anuradha G, M.Vani Pujitha “House Price Prediction Using Regression Techniques: A Comparative Study”, IEEE 6th International Conference on smart structures and systems ICSSS 2019
3. Feng Wang, Ynag Zoe, Haoyu Zhang and Haodang Shi.“House price prediction approach based on Deep Learning and ARIMA model”. 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT).
4. Zhongyun, Jiang, Guoxin, Shen, “Prediction of House Price Based on The Back Propagation Neural Network in The Keras Deep Learning Framework”, 2019 6th International Conference on Systems and Informatics (ICSA I2019).
5. Aboozar Taherkhani , Georgina Cosma , T.M. McGinnity, “AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multiclass imbalanced datasets using transfer learning”, Neurocomputing 2020
6. Zhen Peng, Qiang Huang, Yincheng Han, “Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm”, 2019 IEEE 11th International Conference on Advanced Infocomm Technology.
7. Cun Jia, Xiunan Zoua, Yupeng Hub, Shijun Liub, Lei Lyua, Xiangwei Zhenga, “XG-SF: An XGBoost Classifier Based on Shapelet Features for Time Series Classification”, Procedia Computer Science 147 (2019) 24–28.
8. Qiansheng Zhang, Ziqi Wu, Qiting Chen & Lushuo Wei, “Fuzzy Prediction Method For Stock Price Based On An Enhanced Boosting Algorithm”, October-December 2017.
9. Wu, Jiao Yang, "Housing Price prediction Using Support Vector Regression", Spring 5-31-2017.
10. Dr. Swapna Borde1, Aniket Rane2, Gautam Shende3, Sampath Shetty4,”Real Estate Investment Advising Using Machine Learning”, Mar -2017.
11. Ahmed Khalafallah ,Neural Network Based Model for Predicting Housing Market Performance, Tsinghua Science & Technology 13(S1):325-328 ,October 2008.
12. Nihar Bhagat, Ankit Mohokar, Shreyash House Price Forecasting using Data Mining.International Journal of Computer Applications 152(2):23-26, October 2016.
13. Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy,Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study: Malang, East Java, Indonesia. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017, Page(s):323-326
14. Model,Azme Bin Khamis, Nur Khalidah Khalilah Binti Kamarudin, Comparative Study On Estimate House Price Using Statistical And Neural Network, International journal of scientific and technology,research volume 3, ISSUE 12, December 2014,Page(s):126-131.
15. Steven C. Bourassa, Eva Cantoni, Martin Edward Ralph Hoesli,Spatial Dependence, Housing Submarkets and House Price Prediction The Journal of Real Estate Finance and Economics, 143-160, 2007.
16. Nihar Bhagat, Ankit Mohokar, Shreyash House Price Forecasting using Data Mining. International Journal of Computer Applications 152(2):23-26, October 2016.
17. Ahmed Khalafallah ,Neural Network Based Model for Predicting Housing Market Performance, Tsinghua Science & Technology 13(S1):325-328 ,October 2008.
18. Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy,Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study: Malang, East Java,

Indonesia. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017, Page(s):323-326

19. Q. You, R. Pang, L. Cao and J. Luo, "Image based appraisal of real estate properties," IEEE Transactions on Multimedia, vol. 19, no. 12, pp. 2751–2759, 2017.
20. Kumar, B. Anil, C. U. Anand, S. Aniruddha and U. Kumar, "Machine learning approach to predict real estate prices," Discovery, vol. 44-205, pp. 173-178, 2015.
21. Punithavathani, D. Shalini, K. Sujatha, and J. Mark Jain. "Surveillance of anomaly and misuse in critical networks to counter insider threats using computational intelligence." Cluster Computing 18.1 (2015): 435-451.
22. Sujatha, K., and D. Shalini Punithavathani. "Optimized ensemble decision-based multi-focus imagefusion using binary genetic Grey-Wolf optimizer in camera sensor networks." Multimedia Tools and Applications 77.2 (2018): 1735-1759.

23. Chang, Jinping, Seifedine Nimer Kadry, and Sujatha Krishnamoorthy. "Review and synthesis of Big Data analytics and computing for smart sustainable cities." IET Intelligent Transport Systems (2020).