IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10468

HOUSE PRICE PREDICTION USING MACHINE LEARNING

Khan Sohil Liyaqatullah[1], Usman Malik[2], Qureshi Mohammed Basheeruddin[3]

Student, Department Of Information Technology, M. H. Saboo Siddik College Of Engineering, Mumbai, India[1-3]

Abstract : In today's world, everyone wishes for a house that suits their lifestyle and budget and provides amenities according to their needs. House prices keep on changing very frequently which proves that house prices are often exaggerated. There are many factors that have to be taken into consideration for predicting house prices such as location, number of rooms, carpet area, how old the property is? and other basic local amenities. This research aims to predict house prices based on every basic parameter that is considered while determining the price.

Keywords: Machine learning , Supervised learning , linear regression , model , Ridge regression

I. INTRODUCTION :

Data is at the heart of technical innovations, achieving any result is now possible using predictive models. Machine learning is extensively used in this approach. Machine learning means providing valid dataset and further on predictions are based on that, the machine itself learns how much importance a particular event may have on the entire system based on its pre-loaded data and accordingly predicts the result. Various modern applications of this technique include predicting stock prices, predicting the possibility of an earthquake, predicting company sales and the list has endless possibilities . For our research project, we have considered Bengaluru as our primary location and are predicting real-time house prices for various localities in and around Bengaluru. We have used parameters like 'square feet area', 'no. of Bedrooms', 'No of Bathrooms', etc. We have taken into account a verified dataset with diversity so as give accurate results for all conditions and develop a real estate valuation model which predicts the value of a property using the domain of Machine Learning. The algorithmic approach involves usage ridge regression on top of linear regression approach (Supervised Learning). We use various regression techniques in this pathway, and our results are not sole determination of one technique rather it is the weighted mean of various techniques to give most accurate results. The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied.

II. SCOPE & OBJECTIVES

This new model will help the new purchasers and less experienced clients to comprehend the pace of the property that are over-appraised or under-evaluated. Presently, the cost of the property rely upon parameters of the land in the monetary framework and the public.We have thought about different basic parameters, (for example, number of rooms, living zone and so forth).

At that point these parameter esteems are applied in Linear Regressor model calculations. We have estimated direct linear regression is applied to anticipate the selling pace of an entity

In this methodology we are foreseeing house value esteems utilizing Linear relapse with edge regularization way to deal with decline the blunder inactivity and furthermore for examination dependent on different mistake measurements, for example, Mean Absolute Error (MAE), Mean Squared Error (MSE), R- Squared worth and Root Mean Squared Error (RMSE).

In Supervised learning, the algorithm consists of a target variable or a dependent variable which is to be predicted from a set of independent variables. Using a function, the inputs are mapped to the desired outputs.

The real estate industry has become a competitive and nontransparent industry. The data mining process in such an industry provides an advantage to the developers by processing those data, forecasting future trends and thus assisting them to make favorable knowledge-driven decisions.

Our main focus here is to develop a model which predicts the property cost for a customer according to his\her interests. Our model analyses a set of parameters selected by the customer so as to find an ideal price according to their requirements and interest. It uses a classical technique called linear regression, forest regression and Boosted regression for prediction and tries to give an analysis of the results obtained. On top of this, Neural networks are further used to increase the accuracy of the algorithm which are then further enhanced with boosted regression. It helps establishes the relationship strength between dependent variable and other changing independent variable known as label attribute and regular attribute respectively.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10468

III. LITERATURE SURVEY

Housing prices indicate the current economic situation and also are a concern to the buyers and sellers. There are many factors that have an impact on house prices, such as the number of bedrooms and bathrooms, House price depends upon its location as well. Predicting house prices manually is a difficult task and generally not very accurate, hence there are many systems developed for house price prediction.
Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh had proposed an advanced house prediction system using linear regression. This system's aim was to make a model that can give us a good house price prediction based on other variables. They used the Linear Regression for Ames dataset and hence it gave good accuracy. The house price prediction project had two modules namely, Admin and the User. Admin can add location and view the location. Admin had the authority to add density on the basis of per unit area. Users can view the location and see the predicted housing price for that particular location. housing prices indicate the current economic situation and also are a concern to the buyers and sellers.

There are many factors that have an impact on house prices, such as the number of bedrooms and bathrooms, House price depends upon its location as well. Predicting house prices manually is a difficult task and generally not very accurate, hence there are many systems developed for house price prediction.

This paper [1] proposed on Hybrid Regression technique for housing Prices Prediction focused on the use of creative feature engineering to find the optimal features and their correlation with Sales Prices. Feature engineering improved the data normality and linearity of data. Their system showed that working on the Ames Housing dataset was convenient and showed that the use of Hybrid algorithms (65% Lasso and 35% Gradient Boost) provided results in predicting the house prices rather than using one from lasso, ridge or gradient boost.
There are several factors that affect house prices .In his research Rahadi, et al. divide these factors into three main groups,including the size of the house, the number of bedrooms, the availability of kitchen , the availability of the garden, the area of land and buildings, and the age of the house while the concept is an idea offered by developers who can attract potential buyers, for example, the concept of a minimalist home, healthy and green environment, and elite environment.

IV. PROPOSED SYSTEM

The world is shifting from manual to automated systems. The objective of our project is to reduce the problems faced by the customer. In the present situation, the customer visits a real estate agent so that he/she can suggest suitable showplaces for his investments. But the above method is risky as the agent may forecast wrong prices to the customer and that will lead to loss of customer's investment. This manual technique which is currently used in the market is outdated and has a high risk. So as to overcome the drawback, there is a need for an updated and automated system. In our proposed system, the initial step is data scraping. It is a technique with the help of which structured data can be extracted from the web or any application and saved to a database or spreadsheet or CSV file. We will be using the UIPath Studio Platform to develop our RPA Flowchart. UiPath studio also provides the power data scraping with the assistance of scraping wizards. After Data Extraction, we perform Data Cleaning. It refers to the modifications applied to the data before feeding it to the algorithm. Data Cleaning is a technique that is used to convert the raw data into a clean data set wherein we deal with missing data, categorical data as per the required needs. We have cleaned up our entire dataset and also truncated the outlier values. After completion of Cleaning, we will apply variousalgorithms.
There are many algorithms that can be used to predict the house rate. Ridge Regularization, Multiple Linear Regression , Random Forest are some of the algorithms that can be used. We will be using these algorithms for the prediction: Random Forest is a trademarked term for an ensemble of decision trees. In Random Forest, we have many decision trees. Each tree gives a classification to classify a new object based on the attributes which mean that the tree votes for that class. The forest chooses the classification having the most votes (over all the trees within the forest). In short, with Random Forest we can train the model efficiently for small amounts of data and can get pretty good results. It will, however quickly reach some extent where more samples won't improve the accuracy.

Algoirthms used:

1. Linear Regression: Linear regression is the most simple method for prediction. It uses two things as variables which are the predictor variable and the variable which is the most crucial one first whether the predictor variable These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The equation of the regression equation with one dependent and one independent variable 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.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10468

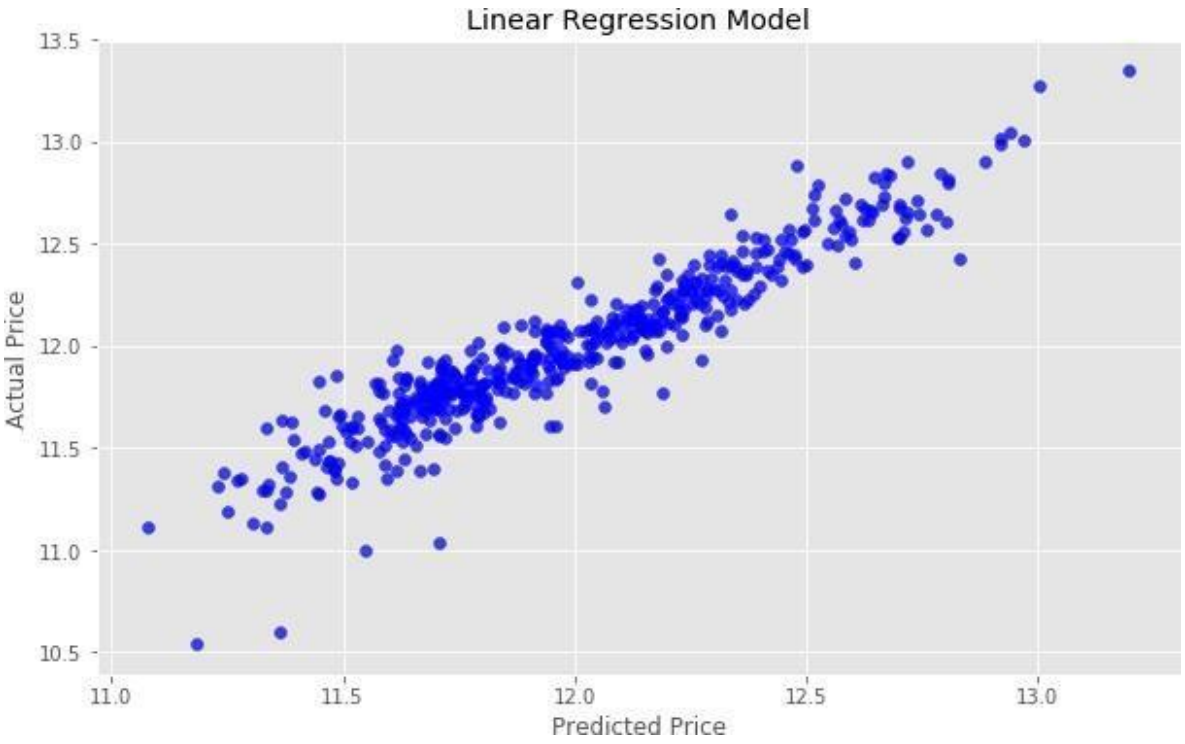


Fig. Linear Regression Scatter Plot

2. Forest Regression : Forest regression uses the technique called as Bagging of trees. The main idea here is to decorrelate the several trees. We then reduce the Variance in the Trees by averaging them. Using this approach, a large number of decision trees are created [3]. Random forest training algorithm applies the technique of bootstrap aggregating, or bagging, to tree learners[7]. Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, 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 Xb, Yb. 2. Train a classification or regression tree fb on Xb, Y

Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, 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 Xb, Yb. 2. Train a classification or regression tree fb on Xb, Yb
2. Train a classification or regression tree fb on Xb, Yb

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

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

individual regression trees on a':

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':

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

IJARCCE

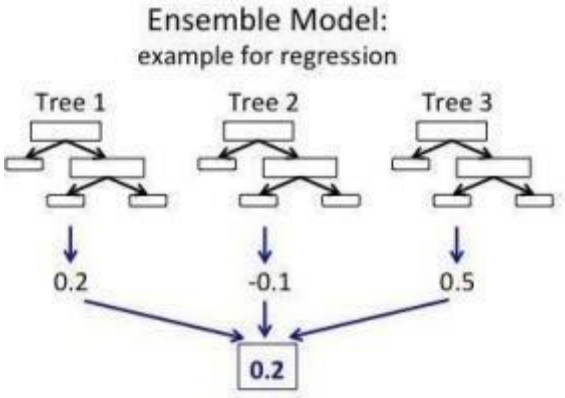
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10468

Representation of it is as follows:



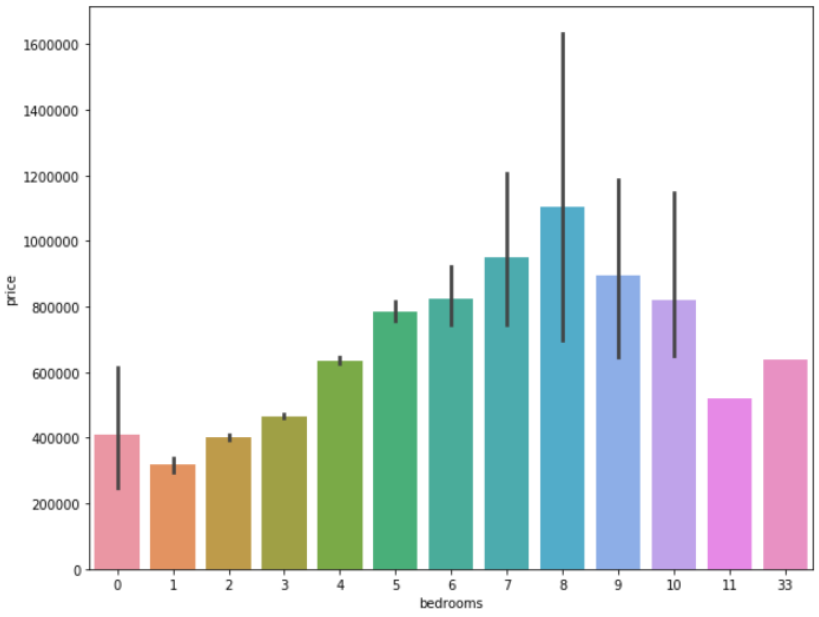


Figure. Data Distribution

The above figure shows the comparison between Overall quality and Salesprice

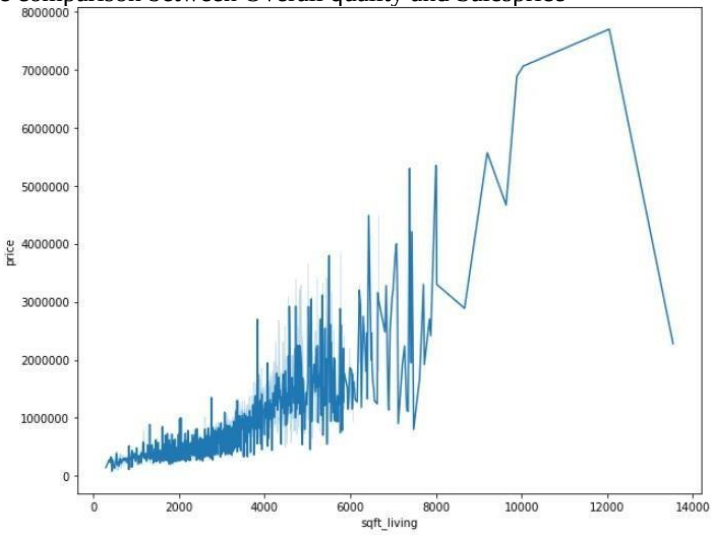


Figure 3. Histogram

IJARCCE

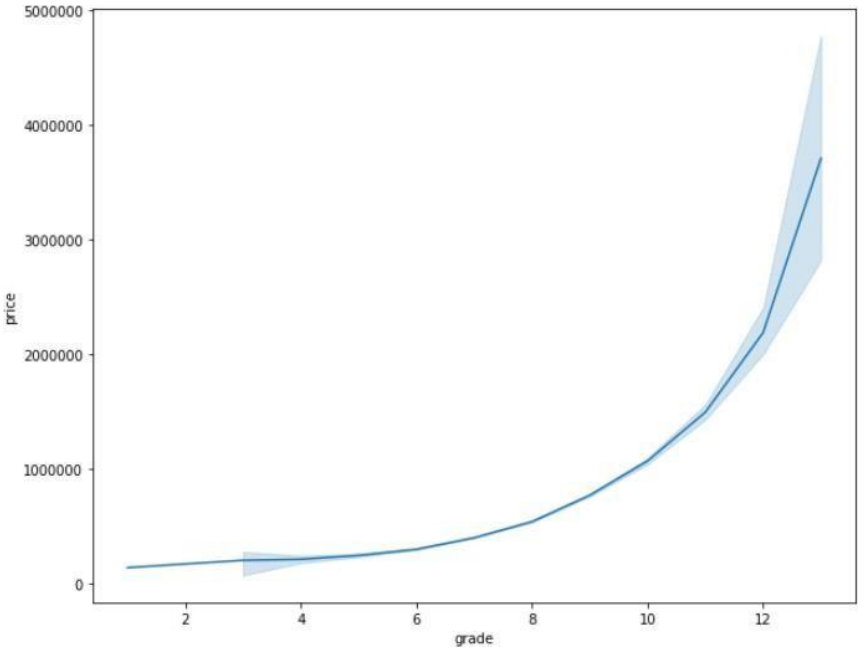
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10468

It shows the validating distributing of data (Salesprice).The above figure shows the distribution of Salesprice value in the dataset



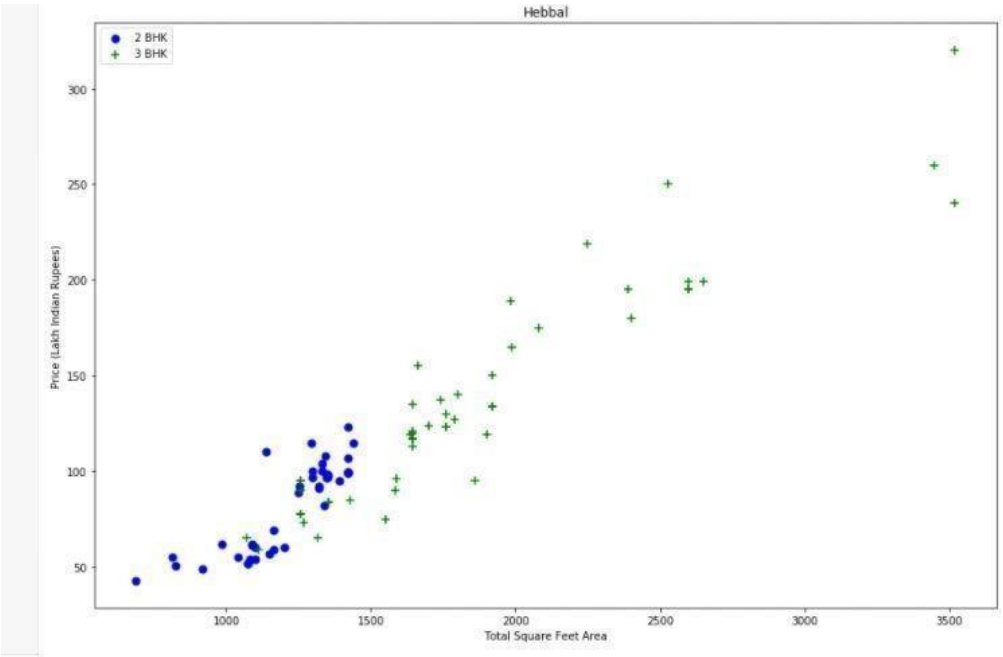


Fig. Data Plot

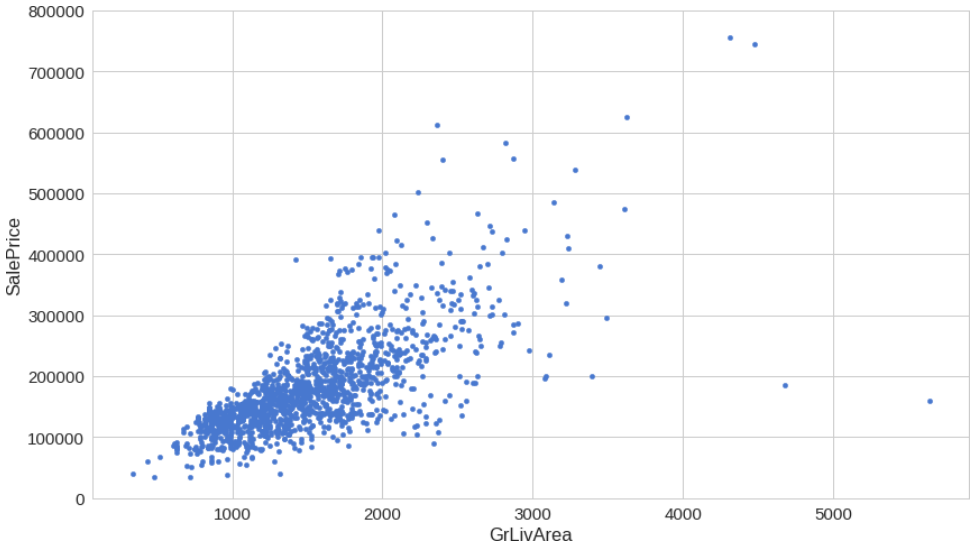
IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

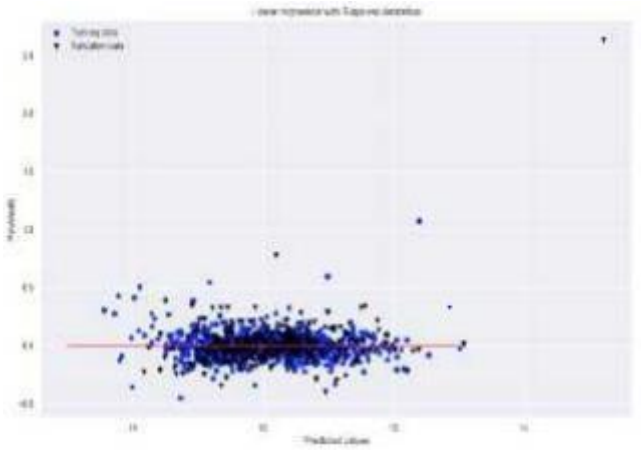
International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

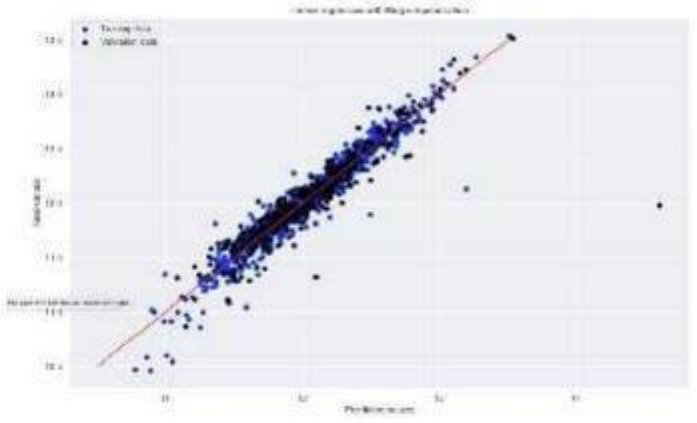
DOI 10.17148/IJARCCE.2021.10468



It is the comparison of predicted values versus residuals. The above outline shows the anticipated qualities change with residuals on y-pivot



Comparison of predicted values versus residuals (After Ridge regression). The above outline shows the anticipated qualities difference with residuals on y-pivot.



Comparison of predicted vs Real values. The above figure indicates the real values vs predicted values applied across the test and training data

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10468

V. METHODOLGY :

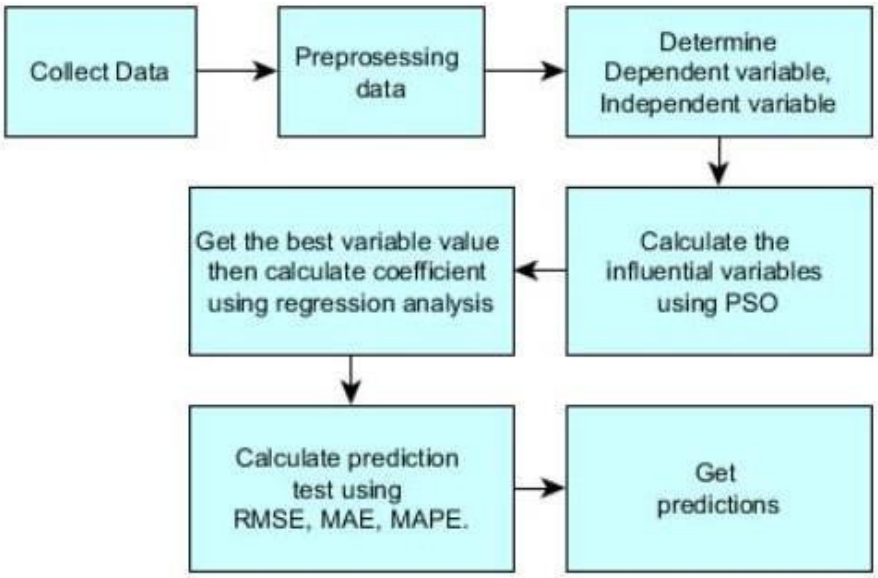


Fig. 1. Diagram flow research.

The selling price is estimates using by considering various parameters such as population rate in particular area, distance to roadways, property age etc. The dataset collection is taken from a standard source such that 80 parameters along with 1000's of test and training data are considered for property valuation and separate dataset is considered for testing and training a model. For further improvement of accuracy, Ridge regularization is applied on top of linear regression so that data are regularized with increase in model accuracy. Users who are going to sell the property can get the accurate values based on this regression prediction. Users requires no intermediate person (broker) to sell in the entity. The python language with its standard libraries are utilized for model expectations dependent on dataset esteem. Since end- user can't run this model each and every time by utilizing python idle there comes the usability lab. To overcome this as well as for powerful utilization of this model by end-users a separate site page is structured with the goal that clients can legitimately pass esteems from site to python code and get the exact value for the entity

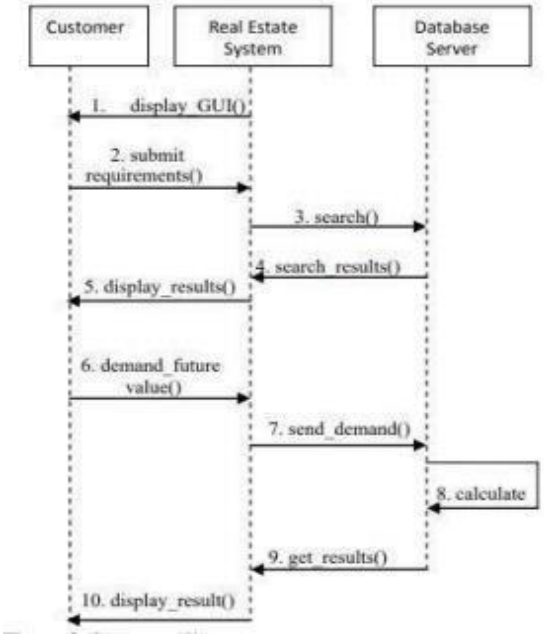


Fig. Sequence Diagram

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10468

The sequence diagram above displays the working of the system. There are 3 objects namely: customer, database and web interface. This also includes the computational mechanisms described in the algorithm. The customer is displayed with the GUI where he can enter the locality, area and different parameters about the house he is looking to buy. The system then displays the matching properties and its price according to the user preferences .

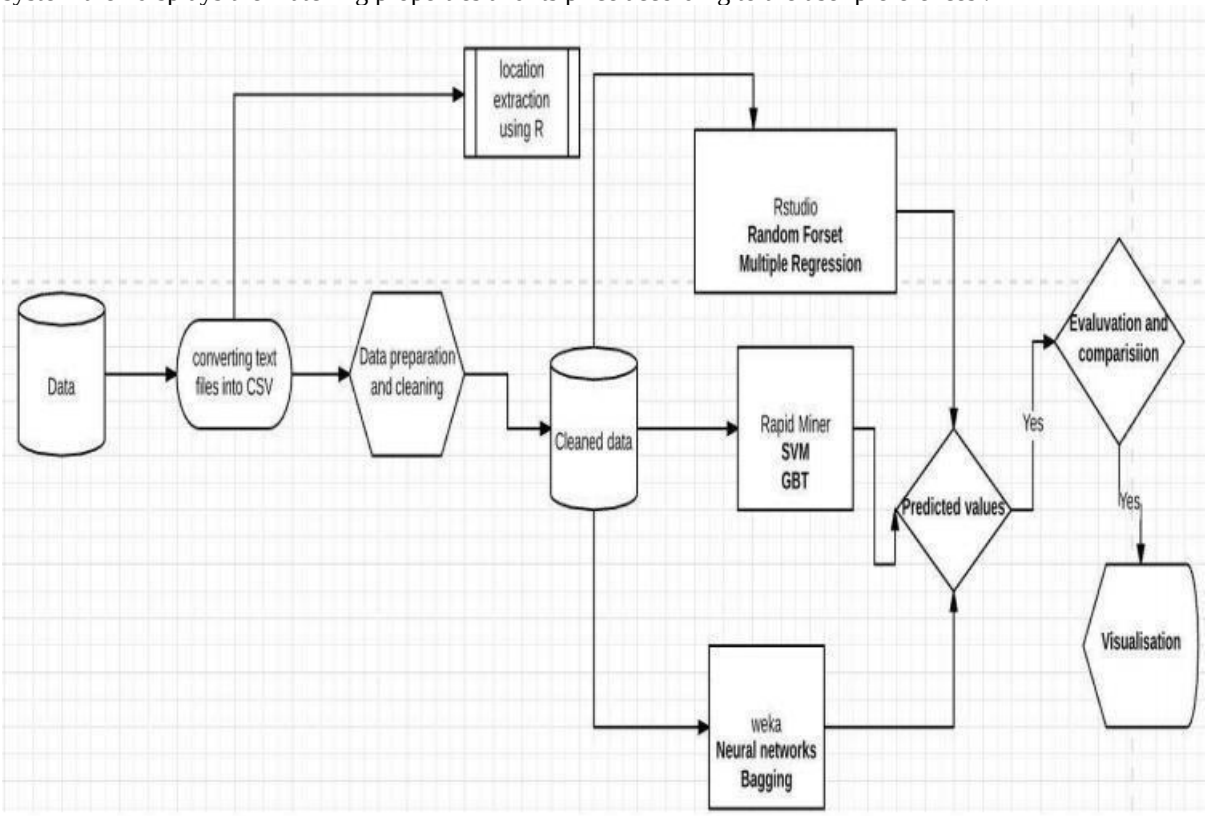


Fig. Architecture of the system

This new model will help the new purchasers and less experienced clients tocomprehend the pace of the property that are over-appraised or under-evaluated. Presently, the cost of theproperty rely upon parameters of the land in the monetary framework and the public.We have thought about different basic parameters, (for example, number of rooms, living zone and so forth). At that point these parameter esteems are applied in Linear Regressor model calculations. We have estimated direct linear regression is applied to anticipate the selling pace of an entity.In this methodology we are foreseeing house value esteems utilizing Linear relapse with edge regularization way to deal with decline the blunder inactivity and furthermore for examination dependent on different mistake measurements, for example, Mean Absolute Error (MAE), Mean Squared Error (MSE), R- Squared worth and Root Mean Squared Error (RMSE).

In Supervised learning, the algorithm consists of a target variable or a dependent variable which is to be predicted from a set of independent variables. Using a function, the inputs are mapped to the desired outputs.

To develop a real estate valuation model which predicts the value of a property using the domain of Machine Learning. The algorithmic approach involves usage ridge regression on top of linear regression approach (Supervised Learning). We use various regression techniques in this pathway, and our results are not sole determination of one technique rather it is the weighted mean of various techniques to give most accurate results. The results proved that this approach yields minimum error and maximum accuracy than individual algorithms applied.

The selling price is estimates using by considering various parameters such as population rate in particular area, distance to roadways, property age etc. The dataset collection is taken from a standard source such that 80 parameters along with 1000's of test and training data are considered for property valuation and separate dataset is considered for testing and training a model. For further improvement of accuracy, Ridge regularization is applied on top of linear regression so that data are regularized with increase in model accuracy. Users who are going to sell the property can get the accurate values based on this regression prediction. Users requires no intermediate person (broker) to sell in the entity. The python language with its standard libraries are utilized for model expectations dependent on dataset esteem. Since end- user can't run this model each and every time by utilizing python idle there comes the usability lab. To overcome this as well as for powerful utilization of this model by end-users a separate site page is structured with the goal that clients can legitimately pass esteems from site to python code and get the exact value for the entity.

IJARCCE

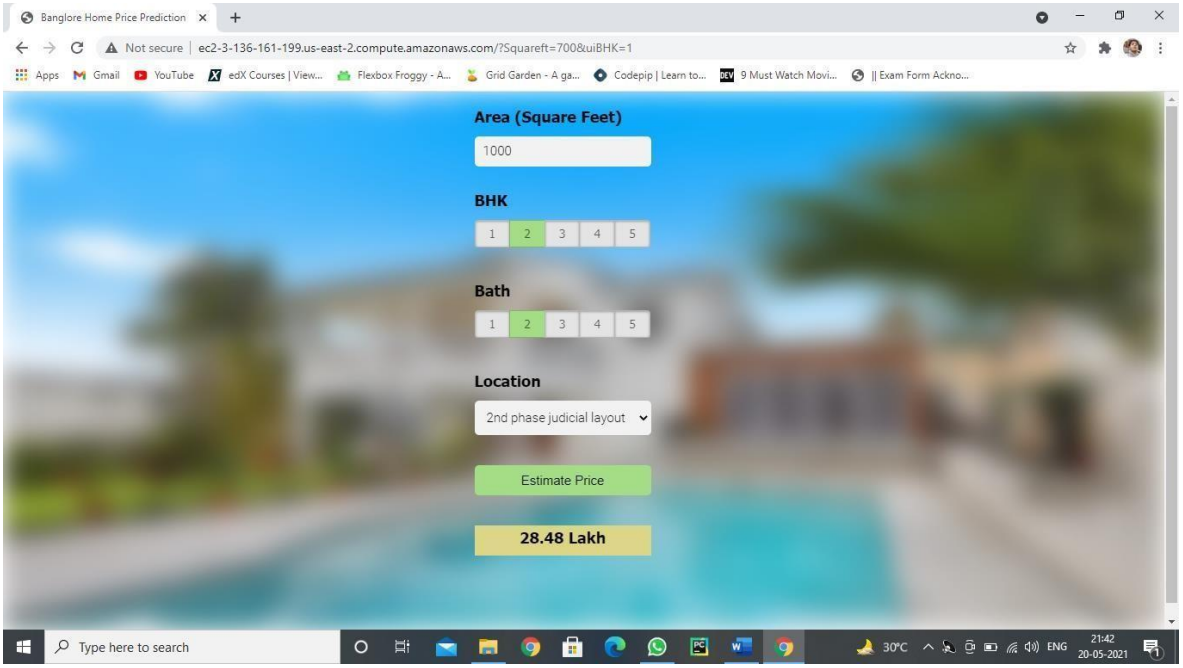
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10468

SCREENSHOT :



VI. CONCLUSION

A system that aims to provide a reliable prediction of housing prices based on test data has been developed. The system makes use of both Linear Regression and Ridge Regularization. The system will get the user parameter values directly from webpage and projects the output based on the trained data. The system will satisfy customers by providing accurate output and preventing the risk of investing in the wrong house. Additional features for the customer's benefit can also be added to the system without disturbing it's core functionality. A major future update could be the addition of larger cities to the database, which will allow our users to explore more houses and which will permit the users to investigate more house datasets, commercial places and to get more precision and consequently go to an appropriate choice

VII. REFERENCES

[1] A. Adair, J. Berry, W. McGreal, Hedonic modeling, housing submarkets and residential valuation, Journal of Property Research, 13 (1996) 67-83.
[2] O. Bin, A prediction comparison of housing sales prices by parametric versus semi- parametric regressions, Journal of Housing Economics, 13 (2004) 68-84.
[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. [Online].
[5] T. Kauko, P. Hooimeijer, J. Hakfoort, Capturing housing market segmentation: An alternative approach based on neural network modeling, Housing Studies, 17 (2002)875-894.
[6] R. J. Shiller, "Understanding recent trends in house prices and home ownership," National Bureau of Economic Research, Working Paper 13553, Oct. 2007. DOI: 10.3386/w13553. [Online].
[7] The elements of statistical learning, Trevor Hastie - Random Forest Generation
[8] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
[9] S. Yin, S. Ding, X. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," IEEE Transactions on Industrial Electronics, 2014.
[10] Friedman, J. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29(5):1189–1232.
[11] . T. Azuma et al., "A survey of augmented reality," Presence, vol. 6, no. 4, pp. 355–385, 1997