House Price Prediction Using Regression Techniques: A Comparative Study

[1]CH.Raga Madhuri, [2] Anuradha G, [3] M.Vani Pujitha
[1,3]Assistant Professor, [2]Associate Professor
[1,2,3]Department of CSE, VRSEC, Vijayawada
[1]chragamadhuri@vrsiddhartha.ac.in, [2]ganuradha@vrsiddhartha.ac.in, [3]pujitha.vani@gmail.com

Abstract-People are careful when they are trying to buy a new house with their budgets and market strategies. The objective of the paper is to forecast the coherent house prices for non-house holders based on their financial provisions and their aspirations. By analyzing the foregoing merchandise, fare ranges and also forewarns developments, speculated prices will be estimated. The paper involves predictions using different Regression techniques like Multiple linear, Ridge, LASSO, Elastic Net, Gradient boosting and Ada Boost Regression. House price prediction on a data set has been done by using all the above mentioned techniques to find out the best among them. The motive of this paper is to help the seller to estimate the selling cost of a house perfectly and to help people to predict the exact time slap to accumulate a house. Some of the related factors that impact the cost were also taken into considerations such as physical conditions, concept and location etc.

Keywords - [DM]Datamining, non-householders, [p]prediction, regression, landholdings, location.

I. INTRODUCTION

This article refers together with latest Forecast on Research predictions considering trends to further plan their economics. The main motivation of the project FORECASTING VARIATIONS ON HOUSE PRICE was to make the best possible prediction of house prices by using appropriate algorithms and finding out which among them is best suitable for predicting the price with low error rate. This is an interesting problem because most of the people will eventually buy/sell a home. This problem allows us, as house price analysts, to learn more about the housing market and helps with making more informed decisions. The analysis that were done in this paper is mainly based on the datasets of Vijayawada, A.P. because of unexpected changes in price of houses in and around Vijayawada due to emergence of new capital city Amaravati because of formation of new state.

In this paper ,we try to demonstrate all the possible Regression techniques which are suitable to our problem. The brief overview of all the reference taken are as follows: In [1] [MLR]Multiple Linear Regression is used which uses more than one attributes for prediction and in [2],[9] [RR]Ridge and[LR] LASSO Regressions are used in which Ridge regression regularizes the [rc] regression coefficient by posing a interest on the size.

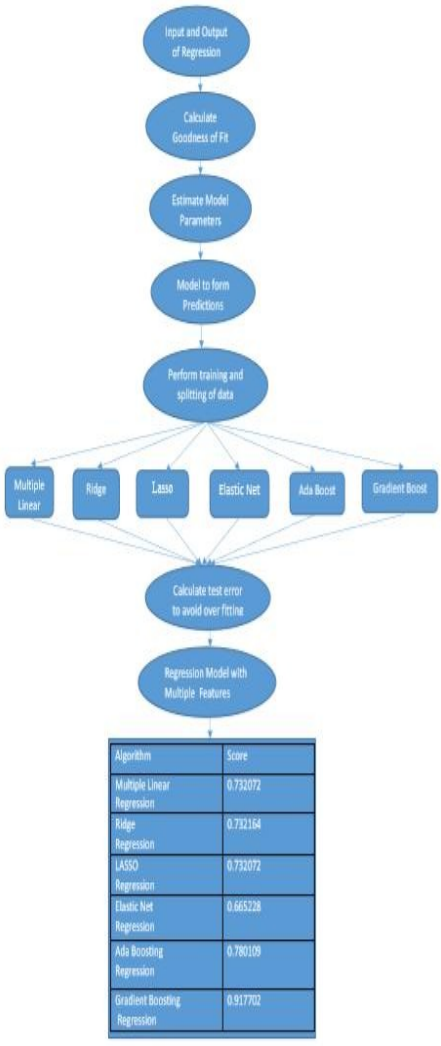

Fig. 1. Data flow Diagram

[LR]LASSO Regression is also same to Ridge but with a little difference, it uses the L1 penalty. In [3] the [ER]Elastic Net Regression was used as a penalization method. In [4] the

classification algorithm Naive Bayes is used. In [5] one of the most efficient Regressions i.e., [GBR]Gradient Boosting Regression is used. In [6],[10],[13] the Artificial Neural Network theory is used. Hedonic Pricing theory is used in [7],[11] which assumes the property value as the sum of its attribute values. Linear Regression model is used in [8][12].In [14][16] Geographically Weighted Regression is used which allows local variations in rate. In [15][17] Bayesian Linear Regression is used. The Data Flow Diagram depicts the order of steps i.e.,the flow in which the research had completed.We try to model predictions by performing training and splitting of data and predict it using various machine learning techniques like different forms of Regression. We have analysed each regression technique and calculated its score.

Now we implement techniques such as [MLR]Multiple Linear Regression,[RR]Ridge Regression, Lasso Regression[LR],[ER] Elastic Net Regression, [AR] Adaboost Regression and [GBR]Gradient Boosting Regression using tools like python programming, Ipython jupyter Notebook GraphLab,Sframes.This work is implemented using Jupyter, which is out of software called IPython Project.The above Figure 1 is used here to represent the flow of data and its processing involved with different regression techniques.

II. METHODOLOGY

Multiple Linear Regression [MLR]Multiple linear regression is the most common form of linear regression .As a forecast predictor, the [MLR]multiple linear regression is used to illustrate the co- relation between continuous dependent variable and two or many independent variables as follows with Eq. 1 and Eq. 2:

$$E(Y|X) = \alpha 1 + \beta 1X1 + \cdots + \beta pXp \quad (1)$$

where α1 is called the intercept as well βj are called coefficients/slopes. If it is further one step, we can resemble how the response(s) keep varying around their respective mean values. The above statement leads to a model as follows:

$$Yj = \alpha 1 + \beta 1Xj,1 + \cdots + \beta pXj,p + j \quad (2)$$

which is yet equivalent to the below statement

$Yj = E(Y|Xj) + j$

Where Yj is actual value and I is error rate. We write Xa,b for the b th predictor variable measured for the a th observation.

[RR] Ridge Regression

[RR]Ridge Regression is a tool for analysis of [MR]multiple regression on the data that have multicollinearity (mcl). Multicollinearity(mcl) is existence of near–linear relationships among the variables which are independent. [RR]Ridge regression applies a special type of condition on parameters as in Eq. 3 and Eq. 4

(ββ's): ^βridgeβ^ridge was chosen to reduce the error of sum of the squares

$$\sum a=1(ya-p\sum b=1xab\beta b)2+\lambda p\sum b=1\beta 2 \quad (3)$$

which is equivalent minimization as

$$\sum ni=1(yi-\sum pj=1xij\beta j)2\sum i=1n(yi-\sum j=1pxij\beta j)2 \quad (4)$$

subject to, for some c>0c>0,

$\sum pj=1\beta 2j<c\sum j=1p\beta j2<c$, i.e. constraining the sum of the squared coefficients.

[LR] LASSO Regression

[LR] Lasso Regression is one of the type of linear Regression which uses the technique of shrinkage as in Eq. 5.

n is the number of observations.

Y_a is the response at observation a.

X_a is data, vector of p values at observation a.

$\lambda 1$ is positive regularization parameter to one of the value of Lambda.

The parameters β, $\beta 0$ are scalar and vector p. As the value of λ increases, the value of β decreases.

[ER] Elastic Net Regression

This technique solves the problem of regularization. For an $\alpha 1$ strictly between the values 0 and 1, and a non negative $\lambda 1$,[ER] elastic net solves the problem as in Eq. 6:

$$\hat{\beta} = \text{argmin}||q-P\beta||^2 + \lambda_{ii}||\beta||^2 + \lambda_i||\beta||_1 \quad (6)$$

[GBR] Gradient Boosting algorithm [GBR]Gradient boosting is a technique of machine learning to solve the regression and classification related problems. It produces as a result a prediction model which ensembles all the weak prediction model mainly decision trees using the Eq. 7 and Eq. 8.

$$\text{Loss} = \text{MSE} = y_i - y_{ip})^2 \quad (7)$$

Where, y_i= ith target, y_{ip} =ith prediction, L (, p) is Loss function

$$y_i^p = y_{ip} + (\alpha 1 * \sum y_t - y_{tp})^2)/ y_{ip} \quad (8)$$

Which $y_{ip} = y_{ip} - \alpha 1 * 2 * \sum (y_t - y_{tp})$

Where, α1 is learning rate and $\sum y_t - y_{tp})$ is sum of residuals.

[AR]AdaBoosting Regression - [AR]AdaBoost is a regression algorithm which is meant for constructing a “strong” classifier which combines both “simple” and “weak” classifier.

III. IMPLEMENTATION

This research makes use of jupyter IDE.It is an open-source web app that helps us to share as well create documents which have livecode,visualizations,equationsa and text that narrates. It contains tools for data cleaning,transformation of data,

simulation of numeric values,modelling using statistics,visualization of data and machine learning tools. Here we collected house sales related data to estimate the house prices based on real world dataset kingcounty.It is a public output dataset of that specified region in USA.Here we used other tools like GraphLab canvas, SFrames for perfect data visualization. All the above mentioned regression techniques are implemented using the above specified tools. In order to find out the efficient regression technique for prediction, we require certain parameters to perform comparision among the techniques. The parameters chosen for the comparison are Scores of the algorithm, [MSE] Mean Square Error and [RMSE] Root Mean Square Error. The below Table 1 represents the resultant summary of the parameters ,when above techniques are implemented practically.

Table I
Comparison of Algorithms

Algorithm	**Score**	**MSE**	**RMSE**
Multiple Linear Regression	0.732072	391875744 48.88446	197958 51699
Ridge Regression	0.732164	391740496 29.73141	197924 35330
LASSO Regression	0.732072	391875537 34.32263	197958 46466
Elastic Net Regression	0.665228	489642930 85.00798	221278 76781
Ada Boost ing Regr ession	0.7801099	32161481 079.94242	179336 22355
Gradient Boosting Regression	0.9177022	12037006 088.27804	109713 90390

From the above table,we can easily perform comparision of different algorithms clearly to find the best among them.Figure 2 below is used to clearly visualize the performance of various techniques in a graphical format based on their scores.In Figure 2,x axis represents the various regression techniques considered for study and y-axis represents the score values observed.

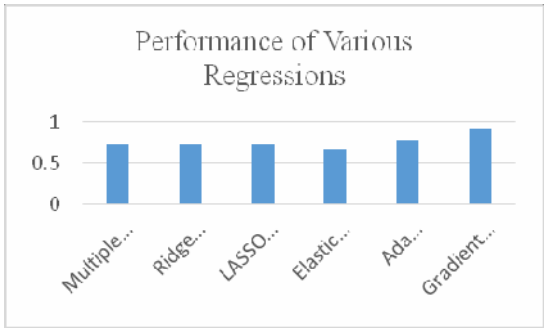

Fig. 2 , Comparison of various Regressions

The graphical representation of all the different regression techniques listed above are clearly represented below using GraphLab canvas .

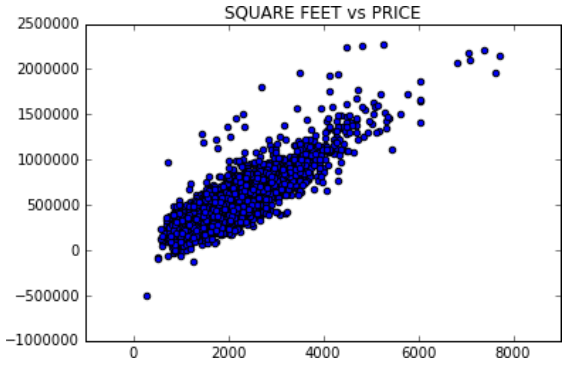

Fig. 3. [MLR] Multiple Linear Regression.

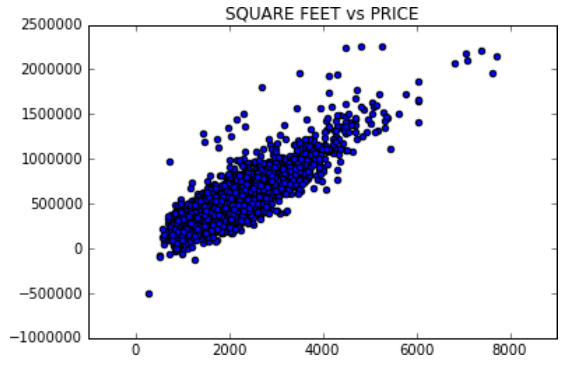

Fig. 4. [RR] Ridge Regression.

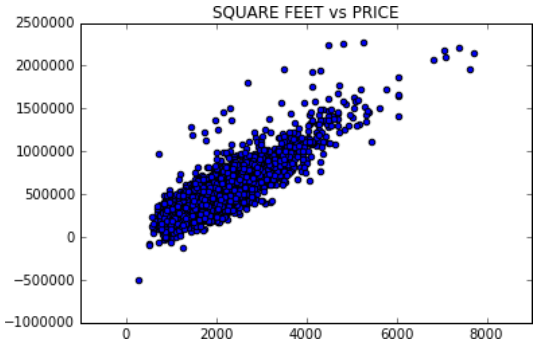

Figure 5: [LR] LASSO Regression.

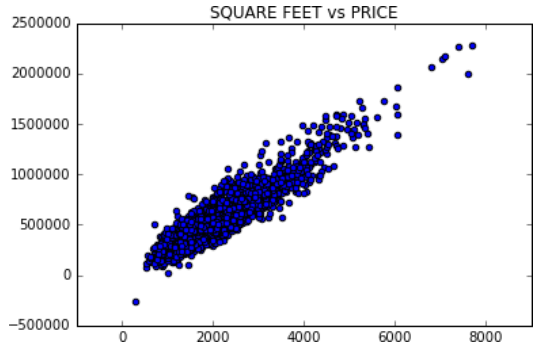

Fig. 6. [ER]Elastic Net Regression.

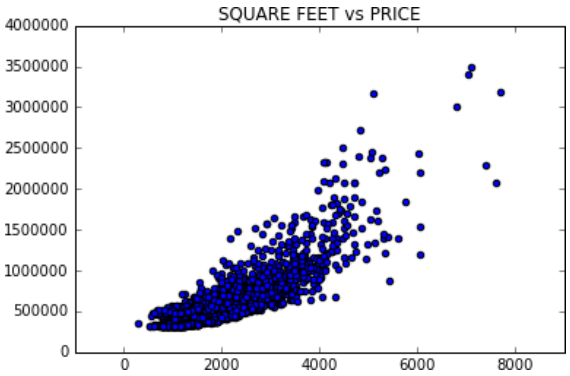

Fig. 7. [AR]Ada Boosting Regression.

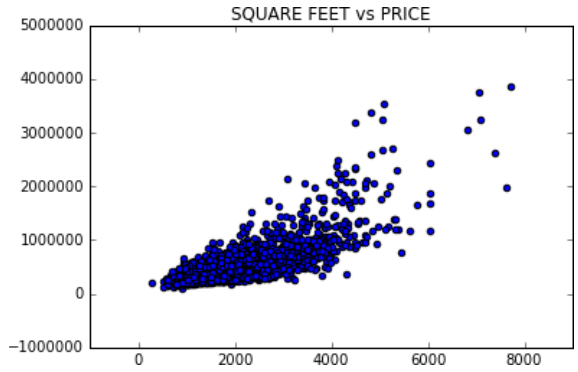

Fig. 8. [GBR]Gradient Boosting Regression.

IV. CONCLUSION

This article mainly concentrates on the comparison between different machine learning algorithms (Multiple Linear Regression, Ridge Regression, LASSO Regression, Elastic Net Regression, Ada Boosting Regression, gradient boosting) about House price prediction Analysis. From the above experiment results, gradient boosting algorithm has high accuracy value when compared to all the other algorithms regarding house price predictions. Here the [MSE] Mean Square Error and [RMSE] Root Mean Square Error are used in order to calculate the accuracy value of the algorithm on the King County Dataset which was collected from public dataset. The paper can be extended by applying the above said algorithms to predict House resale value.

REFERENCES

[1]. Aminah Md Yusof and Syuhaida Ismail ,Multiple Regressions in Analysing House Price Variations. IBIMA Publishing Communications of the IBIMA Vol. 2012 (2012), Article ID 383101, 9 pages DOI: 10.5171/2012.383101.

[2]. Babyak, M. A. What you see may not be what you get: A brief, nontechnical introduction to over fitting regression-type models. Psychosomatic Medicine, 66(3), 411-421.

[3]. Atharva chogle, priyanka khaire,Akshata gaud, Jinal Jain .House Price Forecasting using Data Mining Techniques International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 12, December 2017

[4]. Darshan Sangani,Kelby Erickson, and Mohammad Al Hasan, Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting, IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Page(s):530 - 534

[5]. 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.

[6]. 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

[7]. Nageswara Rao Moparthi,Dr. N. Geenthanjali " Design and implementation of hybrid phase based ensemble technique for defect discovery using SDLC software metrics", An International Conference by IEEE, PP. 268 – 274 ,2016

[8]. Nihar Bhagat, Ankit Mohokar, Shreyash House Price Forecasting using Data Mining.International Journal of Computer Applications 152(2):23-26, October 2016.

[9]. Valeria Fonti ,Feature Selection using LASSO Research Paper in Business Analytics, VU Amsterdam, March 30, 2017.

[10]. Peter B. Luh,Neural Network-Based Market Clearing Price Prediction and Confidence Interval Estimation With an Improved Extended Kalman Filter Method, IEEE Transactions on Power Systems 20(1):59 - 66 , March 2005.

[11]. Visit Limsombunchai, Christopher Gan and Minsoo Lee, House Price Prediction: Hedonic Price Model vs. Artificial Neural Network, Lincoln University, Canterbury 8150, New Zealand, American Journal of Applied Sciences 1 (3): 193-201, 2004.

[12]. Dr. Nageswara Rao Moparthi,,Ch Mukesh, Dr. P. Viday Saga," Water Quality Monitoring System Using IoT", An International Conference by IEEE, PP. 109 – 113 ,2018

[13]. Ahmed Khalafallah ,Neural Network Based Model for Predicting

Housing Market Performance, Tsinghua Science & Technology 13(S1):325-328 ,October 2008.

[14]. 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

[15]. Chris Brunsdon, A.Stewart Fotheringham and Martin E. Charlton, Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity,Geographical analysis, Volume 28, Issue 4,Pages: 281-375,1996.