Journal Of Harmonized Research (JOHR)

Journal Of Harmonized Research in Engineering
6(4), 2018, 38-45

ISSN 2347 – 7393

Original Research Article

PROPOSE ONE ADDED SCHEME IN ADDITION TO LASSO ALONG WITH RIDGE

Mrs. Rajmati[1], Mr. Rakesh Tiwari[2], Mr. Shivendra Dubey [3] Mr. Mukesh Dixit[4]

Radharaman Engineering College, Bhopal, India

Abstract: Regression analysis is an important tool for modeling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized. As given in our paper their there is comparing between Lasso and Ridge, In our work we will introduced one more method and show their comparison, so before implementation let's see some of the factors which are involve during this research.

Keyword- Regression, Lasso, Ridge, SVM, and KNN

Introduction: In insights one of the primary objectives is to construct a model that better speak to a dataset, this procedure incorporate the errand of features selection. The main point of the researcher is to fabricate a model that depicts a reaction variable; keeping in mind the end goal to do as such one of the primary inquiry that the researcher ought to have the capacity to answer is which features/variables would it be advisable for me to think about? Which are the most critical credits to portray the reaction variable? This examination intends to answer this inquiries demonstrating the procedure of feature selection and portraying one of the conceivable methods to achieve this assignment. Specifically the emphasis is on feature selection utilizing the LASSO method.[1]

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

For Correspondence:
vishwas271992@gmail.com
Received on: August 2018
Accepted after revision: September 2018
Downloaded from: www.johronline.com
DOI: 10.30876/JOHR.6.4.2018.38-45

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line). We'll discuss them (figure 1) in detail in the following sections.[2,3]

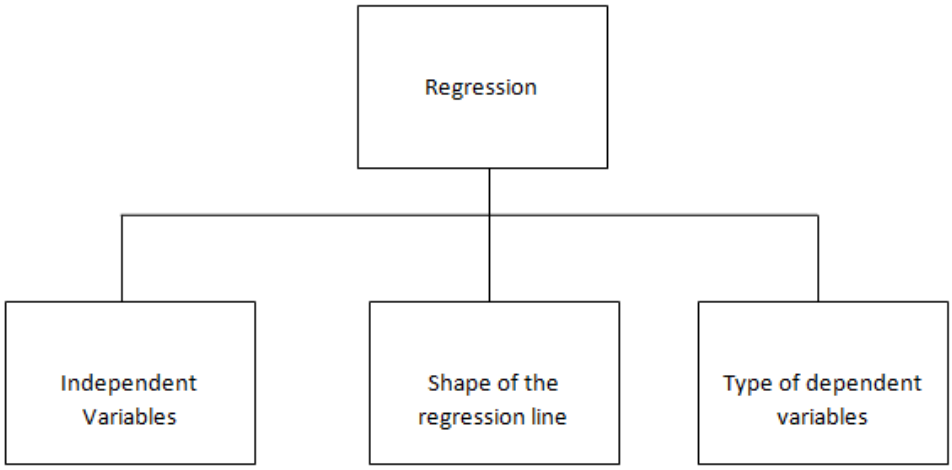

Figure 1: Regression

Regression: Regression is most usually known modeling method. Linear regression is typically amongst the first only some topics which people choose whereas learning predictive modeling. This technique has the dependent variables are continuous, independent variables may be discrete or continuous along with personality of regression lines are linear. Linear Regression found a relationship among dependent variable (B) as well as one or other independent variables (A) via a most excellent fit directly line (also identified as regression line). It's correspond to an equation B = x + y*A + f, where x is intercept and y is slope of the line along with f is error expression. This equation may be use to expect the cost of object variable base on known predictor variables.[12]

Logistic Regression: It is named intended for the function apply at the middle of the scheme, the logistic function. This function is also called as the sigmoid function which was developed through statisticians in the direction of describes property of populace expansion during ecology, rising quickly as well as maxing out at the shipping capacity of the background.[11]

Literature Survey: This examination paper expects to clear up and discuss the utilization of the LASSO strategy toward address the feature selection work. Feature selection is a basic and testing undertaking inside the numerical modeling field, there are a great deal of concentrates that endeavor to upgrade and in addition institutionalize this procedure for some kind of data, yet this is anything but a simple activity. A start of feature selection assignment alongside the LASSO strategy is advertised. We will concern the LASSO feature selection property toward a Linear Regression quandary, and the result of the investigation on a genuine dataset will be uncovered. A similar analysis is dreary on a Generalized Linear Model inside specific a Logistic Regression Model expected for a high-dimensional dataset.[4]

An assortment of estimators are proposed in view of the opening test and Stein-type technique to guess the parameters inside a logistic regression model while it is priori assumed that a few parameters may be obliged to a subspace. Two unordinary punishment estimators since LASSO and additionally ridge regression are likewise estimated. A Monte Carlo replication try was direct for unordinary blends, and the introduction of every estimator was assessed inside terms of recreated near effectiveness. The positive-portion Stein-type shrinkage estimator is proposed for use since its introduction is hearty paying little respect to the

consistency of the subspace information. The arranged estimators are valuable to a genuine dataset to assess their performance.[5]

Linear regression is individual of the broadly utilized statistical techniques available today. It is use by data investigators and also understudies in around each train. Notwithstanding, for the typical ordinary slightest squares technique, there is some extreme presumption finished about data that is as often as possible not valid in certifiable data sets. This can cause a few issues in the littlest sum square model. One of the almost all broad issues is a model overwriting the data. Ridge Regression and also LASSO is two techniques use to improve and extra precise model. I will discuss how overwriting emerge in slightest squares models alongside the thinking for by Ridge Regression and LASSO contain analysis of genuine occurrence data and balance these techniques with OLS and each other to extra construe the advantages and disadvantages of each method.[6]

Regularize regression strategies for linear regression has been created the last just a few decades to crush the imperfections of normal least squares regression through respect to prediction accuracy. In this section, three of these procedures (The Lasso, Ridge regression, and the Elastic Net) are incorporated into CATREG, a best scaling technique for both linear and also nonlinear change of variables inside regression analysis. We clarify that the bizarre CATREG calculation give an exceptionally simple and in addition proficient approach to compute the regression coefficients inside the compelled models expected for the Lasso, Ridge regression, alongside the Elastic Net. The subsequent occasions, subsumed not as much as the expression "regularized nonlinear regression" will be outline by the prostate tumor data, which have before examined in the regularization content expected for linear regression.[7,9]

We think on least - square linear regression quandary with regularization through the one-standard, a situation normally alluded to the same as the Lasso. In this research paper, we show a total asymptotic examination of model steadiness of the Lasso. An assortment of rots of the regularization parameter, we compute asymptotic reciprocals of the probability of precise model selection (i.e., variable decision). For unmistakable rate rot, we exhibit that the Lasso select each variable that must enter the model through probability inclining toward one exponentially quick, while it chooses every one of extra variables with extremely positive probability. We show that this property infer that however we run the Lasso for various bootstrapped replications of a known example, at that point converge the backings of the Lasso bootstrap appraise prompt steady model selection. This novel variable decision calculation, to known as lasso, is contrast positively with promote linear regression techniques lying on manufactured data and additionally datasets as of the UCI machine learning repository.[8,10]

Implementation and its detail: Here we are using information known about a movie in the week of its release, can we predict the total gross revenue for that movie? Such information would be useful to marketers, theater operators, and others in the movie industry, but it is a hard problem, even for human beings. We found that, given a set of numeric, text based and sentiment features from IMDb (Internet Movie database). In this work, we mainly focused on the machine learning regression algorithm. However we have shown the evidences of previous regression algorithm and we have also proposed one regression algorithm which has shown more accuracy than the previous well known algorithms. Here we have implemented the algorithms on the movie_metadata dataset. We have shown the comparison on the basis of train error and test error. We observed that in some condition outperforms than other SVM algorithm. In this work we also included the algorithms like Ridge regression, Bayesian regression, K-NN, Decision tree and SVM. In our proposed work we try to combine the methods of regression algorithm. The dataset contains complete detail about the Hollywood movies including the detail like name, director, release date, facebook likes etc. The whole implementation is done in python 3.6

Proposed Work: In the given base paper we have shown only two regression these are lasso

and Ridge, but in our proposed work we will show more regression and we will show how they perform better than these. Implementation will be in Python.

We collected data from Kaggle related to IMDb for 5043 movies that were released from 1916 to 2016. Our main focus in this research is to find out the movie performance on the basis of imdb score. To evaluate the performance we are using regression analysis so that we can predict whether the movie will perform good or bad. First we have compute the regression analysis on the basis of previous regression algorithms like Ridge regression, Bayesian regression, K-NN, Decision tree and SVM, after that we have proposed our own regression analysis on the basis of previous algorithms. Mainly we had combine the properties of Bayesian and SVM regression analysis algorithm so that they can perform better. The analysis has been done on the previous results.

Regression on movie dataset: During the past 20 years, marketing scholars have developed some prediction models and decision support tools to increase the accuracy of forecast. One mainstream in which is to use multiple linear regression, by making the box office of movie as the dependent variable while the independent variable as the predictors with an impact on box office forecast, to establish a forecast model. It points out some production and marketing characteristic factors influence the financial performance of a movie. That's why we used neural networks in predicting the financial performance of a movie. They compared their prediction model with models that used other statistical techniques; it is found the model built by us do a better job of predicting box office.

Parameters used for analysis: A predictive model is a function which maps a given set of values of the x-columns to the correct corresponding value of the y-column. Finding a function for the given dataset is called training the model. Good models not only avoid errors for x-values they already know, but, in particular, they are also able to create predictions for situations which are only somewhat similar to the situations which are stored in the existing data table. The ability to generalize from known situations to unknown future situations is the reason we call this particular type of model predictive.

4.3. Training Error vs. Test Error

There are two important concepts used in machine learning: the training error and the test error.

Training error. We get this by calculating the classification error of a model on the same data the model was trained on.

Test error. We get this by using two completely disjoint datasets: one to train the model and the other to calculate the classification error. Both datasets need to have values for y. The first dataset is called training data and the second, test data.

5. Algorithm and its flow chart

Let's see how the basic process you can use to calculate the training error for any given dataset and predictive model describe in figure 2:

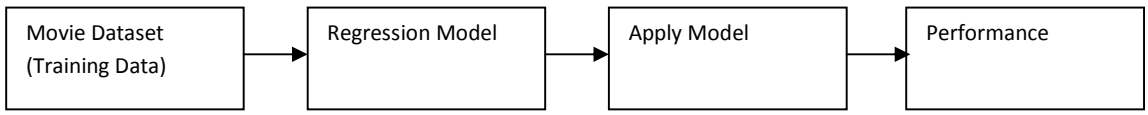

Figure 2: Predictive model for movie data set

First we load a dataset ("Movie_metadata.csv") and deliver this training data into Regression model operator and an "Apply Model" operator which creates the predictions and adds them to the input training data. The last operator on the right, called "Performance," then calculates the training error based on both the true values for y as well as the predictions in p.

Now see (figure 3) the process to calculate the test error. It will soon be apparent why it is so important that the datasets to calculate the test error are completely disjoint (i.e., no data point used in the training data should be part of the test data and vice versa).

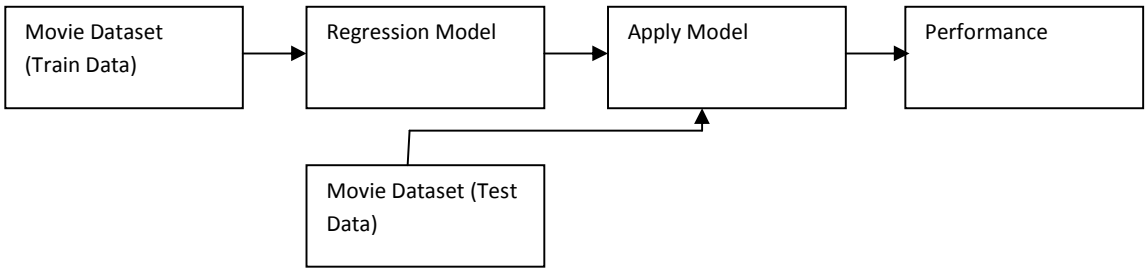

Figure 3: Apply model for calculation of error

Calculating any form of error rate for a predictive model is called model validation. As we discussed, you need to validate your models before they go into production in order to decide if the expected model performance will be good enough for production. But the same model performance is also often used to guide your efforts to optimize the model parameters or select the best model type. It is very important to understand the difference between a training error and a test error. Remember that the training error is calculated by using the same data for training the model and calculating its error rate. For calculating the test error, you are using completely disjoint data sets for both tasks.

Some points which must be considered are as follows:

1. In machine learning, training a predictive model means finding a function which maps a set of values x to a value y
2. We can calculate how well a predictive model is doing by comparing the predicted values with the true values for y
3. If we apply the model to the data it was trained on, we are calculating the training error
4. If we calculate the error on data which was unknown in the training phase, we are calculating the test error

Algorithm:

1. Import all the libraries
2. Initialize f to fetch the data movies_metadata.csv
   f = pd.read_csv("movie_metadata.csv"
   print(data.head)
3. Initialize the train variable
   X_train=data[X_data]
4. Finding normalised array of X_Train through PCA
   pca = PCA().fit(X_std)
5. Initialize number of samples
   number_of_samples = len(y)
6. Computing the train and test error of proposed model

```
error=0
for i in range(len(y_train)):
    error+=(abs(y1_svm[i]-
y_Train[i])/y_Train[i])
train_error_Proposed=error/len(y_Train)*100
print("Train error =
"+'{}'.format(train_error_svm)+" percent"+" in
Proposed Regressor")

error=0
for i in range(len(y_test)):
    error+=(abs(y2_reg[i]-Y_test[i])/Y_test[i])
test_error_Proposed=(error/len(Y_test))*100
print("Test error =
"+'{}'.format(test_error_bay)+" percent"+" in
Proposed Regression")

matplotlib.rcParams['figure.figsize'] = (6.0, 6.0)
preds =
pd.DataFrame({"preds":svm_reg.predict(x_train
), "true":y_train})
preds["residuals"] = preds["true"] -
preds["preds"]
preds.plot(x = "preds", y = "residuals",kind =
"scatter")
plt.title("Residual plot in Proposed")
```

7. Compute the train and test error of the existing models

Result Analysis and its parameters metrics used: In this section we will show how regression algorithms work on the dataset, we will include our own proposed regression algorithm.

Table 1 Result of various regression methods

Regression	Test Error	Train Error
Ridge	14.296076292990001	12.729437097203261
Elastic net	14.274904290676005	14.274904290676005
KNN	5.7683234207599465	12.492260951644399
Bayesian	0.13175302310775863	12.784851827254434
Decision Tree	5.237878057806785	14.264513407018567
SVM	4.0731665386165705	5.772826465611643
Proposed	0.13175302310775863	5.772826465611643

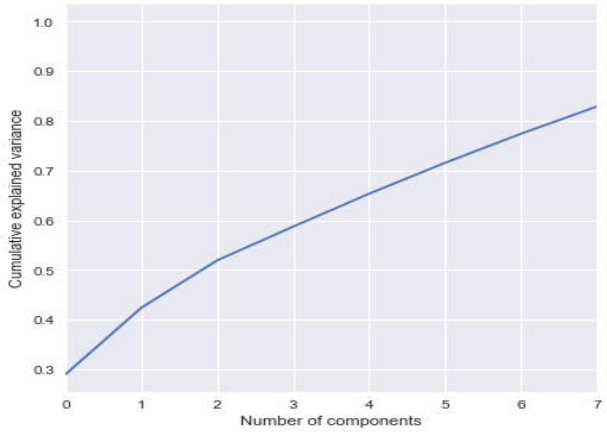

Figure 4: Resultant variance graph

First we will display the result on the basis of test and train error of the previous regression models after that we will display the result of the proposed test and train in table 1 along with all subsequent figures (figure 4 to figure 11) for all regression methods as resultant test error and train error

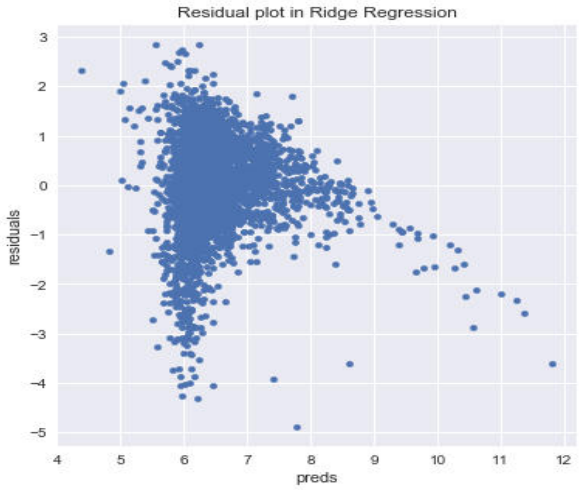

Figure 5: Residual plot in ridge regression

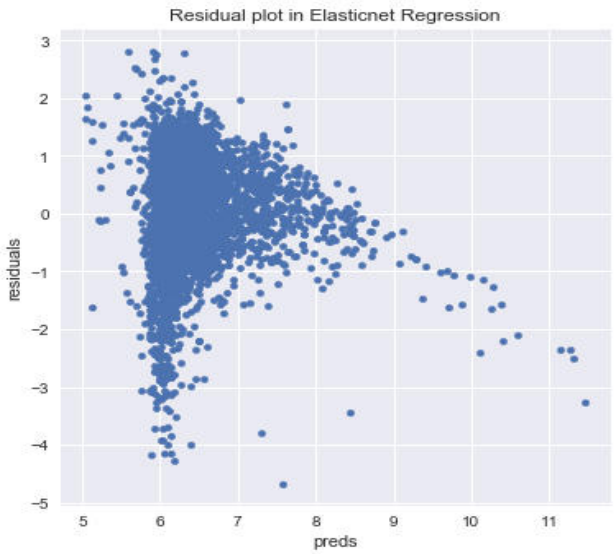

Figure 6: Residual plot in elastic net regression

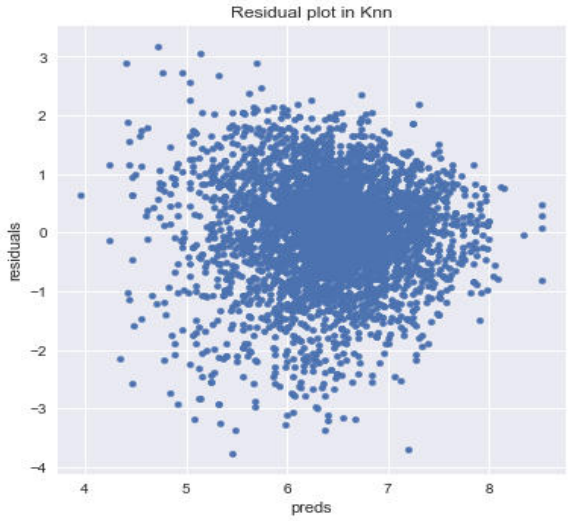

Figure 7: Residual plot in KNN

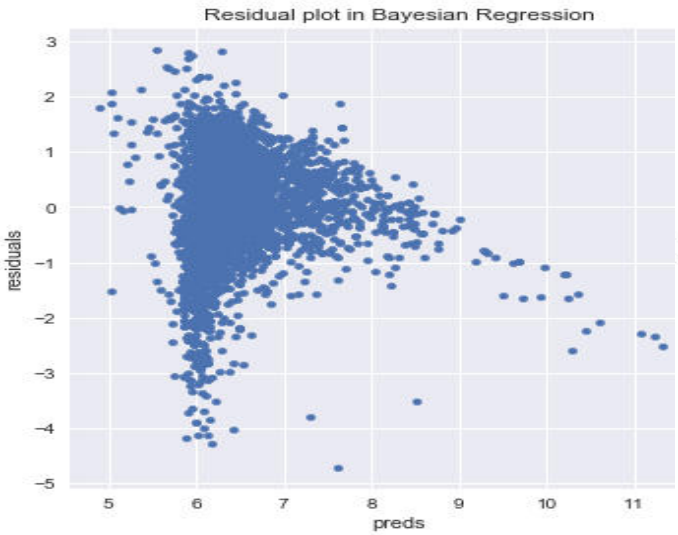

Figure 8: Residual plot in Bayesian Regression

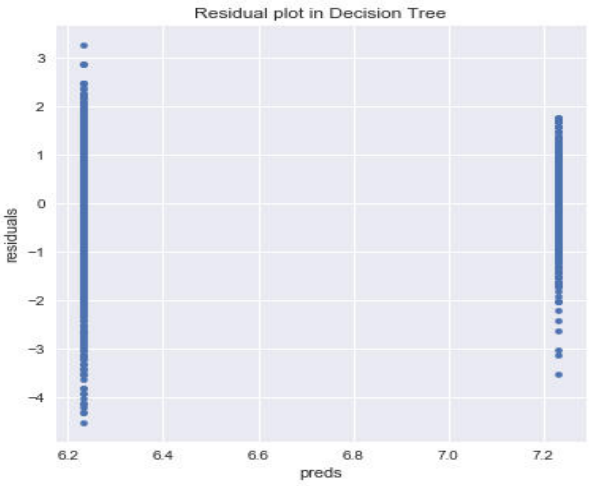

Figure 9: Residual Plot in Decision Tree

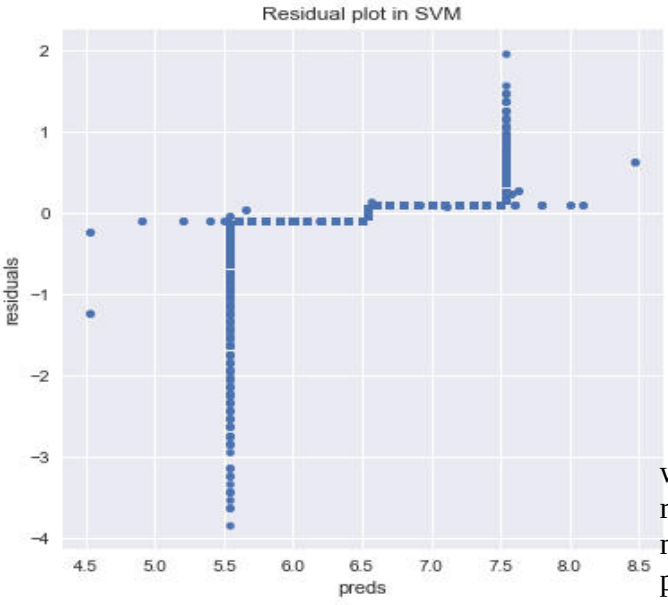

Figure 10: Residual plot in SVM

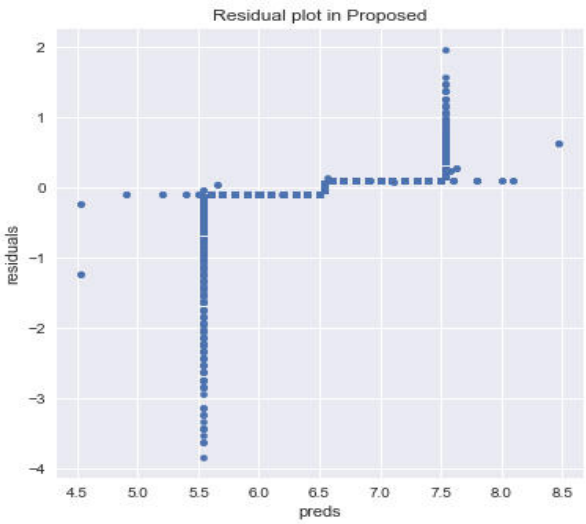

Figure 11: Residual plot in Proposed Regression

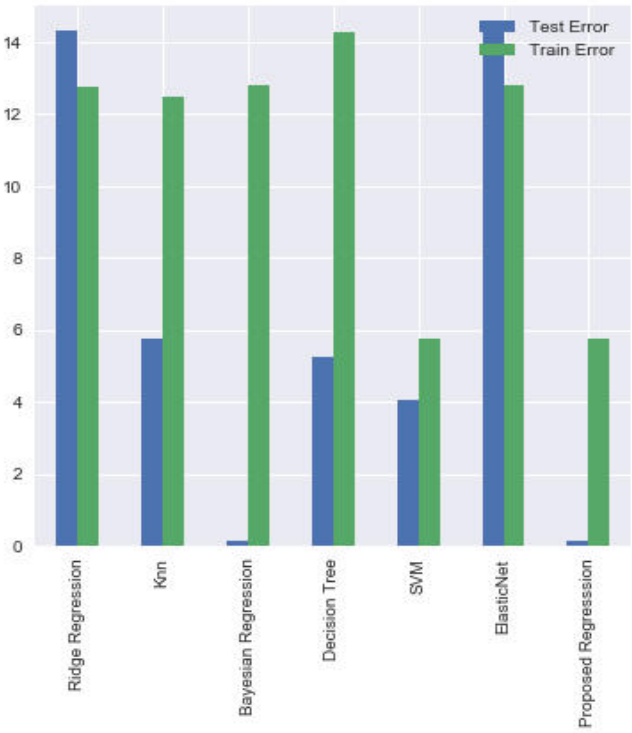

Figure 12: Comparative study of various regression methods

Here, figure 12 describes the comparison of given various regression methods in terms of Test error and Train error.

Conclusion: In the base papers their main focus was on comparison between lasso and ridge regression but in this research we have shown more regression algorithms than the previous papers and even we have proposed one regression algorithm that combine the properties

of SVM and Bayesian. In the proposed algorithm the train and test error are less than other regression algorithm.

References

[1]. L.E. Melkumovaa,, S.Ya. Shatskikh, "Comparing Ridge and LASSO estimators for data analysis", 3rd International Conference "Information Technology and Nanotechnology, ITNT-2017, 25-27 April 2017, Samara, Russia.

[2]. N. Jayanthi , B. Vijaya Babu and N. Sambasiva Rao, "Survey on clinical prediction models for diabetes prediction", Journal of Big Data.

[3]. Jose Manuel Pereira, Mario Basto, Amelia Ferreira da Silva, "The logistic lasso and ridge regression in predicting corporate failure", 3rd GLOBAL CONFERENCE on BUSINESS, ECONOMICS, MANAGEMENT and TOURISM, 26-28 November 2015, Rome, Italy.

[4]. Valeria Fonti, "Feature Selection using LASSO", VU Amsterdam Research Paper in Business Analytics.

[5]. Orawan Reangsephet, Supranee Lisawadi, and Syed Ejaz Ahmed, "A Comparison of Pretest, Stein-Type and Penalty Estimators in Logistic Regression Model", Springer International Publishing AG 2018.

[6]. Chris Van Dusen, "Methods to prevent overwriting and solve ill-posed problems in statistics: Ridge Regression and LASSO", Preprint submitted to Colorado College Department of Mathematics September 16, 2016.

[7]. This chapter has been submitted for publication as Van der Kooij, A.J. & Meulman, J.J. (2006). Regularization with Ridge penalties, the Lasso, and the Elastic Net for Regression with Optimal Scaling Transformations.

[8]. Francis R. Bach, "Bolasso: Model Consistent Lasso Estimation through the Bootstrap", 25 th International Conference on Machine Learning, Helsinki, Finland, 2008.

[9]. Trevor PARK and George CASELLA, "The Bayesian Lasso", Journal of the American Statistical Association June 2008, Vol. 103, No. 482, Theory and Methods.

[10]. Hanzhong Liu and Bin Yu, "Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression", Electronic Journal of Statistics Vol. 7 (2013) 3124–3169 ISSN: 1935-7524.

[11]. Cheolwoo Park and Young Joo Yoon, "Bridge regression: adaptivity and group selection", Department of Statistics, University of Georgia, Athens, GA 30602, USA January 10, 2011.

[12]. Eunho Yang, Aur´elie C. Lozano, Pradeep Ravikumar, "Elementary Estimators for High-Dimensional Linear Regression", 31 st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.