[NC-Rase 18]
DOI: 10.5281/zenodo.1483978

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

IMPLEMENTATION OF REGRESSION TECHNIQUES IN BIG DATA ANALYTICS

Madhura A. Chinchmalatpure[*1] & Dr. Mahendra P. Dhore[2]

[*1]Research Scholar, Department of Electronics and Computer Science, RTM Nagpur University Campus, Nagpur, (MS)-India

[2]Associate Professor, Department of Electronics and Computer Science, RTM Nagpur University Campus, Nagpur, (MS)-India

ABSTRACT

Big data provides collection of large datasets in healthcare field. For analyzing healthcare data, we used Regression as a statistical Technique. It shows the statistical relationship between two or more variables. The statistical technique can be evaluated for the predictive model based on the requirement of the data. This paper deals with different regression techniques applied on large database.So,in this paper, we studied regression techniques wh`ich is compared using the training dataset in order to see correct model for better prediction and accuraracy applied on medicare database.

Keywords: *Big Data Analysis, regression, medicare dataset, Linear, Logistic*

I. INTRODUCTION

In Digitized world, large amount of data is generated, to properly analyze that data big data concept is generated. Big data is the term used to describe collection of large and complex datasets having "4v" definition.volume(amount of data),variety(range of data types and sources), velocity(speed of data in and out) value (e.g.medical images, electronic Health Record (EHR), biometrics data etc.)

The healthcare industry has generated large amount of data so it uses Electronic Health Record(EHR) for patients data, clinical report, doctors prescription, diagnostic reports, medical images, pharmacy information, health insurance related data, data from social media and medical journals. All these information collectively forms big data in healthcare. The medicare database is designed for clinical purposes.

Big data challenges can be divided into

1. Data Challenge: Volume, velocity, variety, veracity, Data Discovery
2. Processing Challenges: Data Collection, Modification of data, Data Analysis, output representation
3. Management Challenges: Data Privacy, Data Security, Governance and ethical issues

Fig. Big Data Challenges

Big data analytics is challenging research area, it refers to tool such as R that are applied to healthcare dataset to obtain the data from database in which to collect current data, preprocess data and analyze data.[1]

Analytics focus on statistical and mathematical analysis of data. The analysis helps to identify the problem from the collected data source. Later it uses various tools and algorithms for better outcomes of data. Regression is supervised

[NC-Rase 18] ISSN 2348 – 8034
DOI: 10.5281/zenodo.1483978 Impact Factor- 5.070
learning. Supervised learning partitions the dataset into training and validation data. There are so many benefits of using regression analysis.[2]

The regression techniques used in this research are applied on large medicare databases are as follows:

- Linear Regression
- Logistic Regression
- Stepwise Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression

In this paper we are applying these techniques on same database as medicare database. These benefits help data analysts to estimate the best set of variables to be used to build the predictive accuracies.

Linear regression analysis is based on a set of input/output observations which are expected to be linearly related. While linear regression methods are highly developed, their applicability to problems with severe uncertainties is still a challenge. Uncertainties may arise from three major sources: (i) measurement uncertainty, inaccuracies in the observations, (ii) model uncertainty. Standard linear regression techniques are based on optimizing a performance criterion, usually the mean squared error, but may fail to provide high, or even acceptable, performance for new observations. When number of the measurement matrix is high, the least squares solution is greatly affected by uncertainties in the measurements and may differ considerably from the underlying linear model[3]

Logistic regression is the logit()—is a regression model where the dependent variable (DV) is categorical. The simplest example of a logit derives from a 2*2 contingency table. In logistic regression binary dependent variable—that is, the output can take only two values, "0" and "1". Logistic regression is useful for testing hypotheses about relationships between a categorical outcome variable and one or more continuous predictor variables.[4]

Stepwise regression is developed as automatic computational procedure, larger number of input variables, the greater benefit it has. Stepwise selection has two main approaches as the forward selection and backward elimination and a combination of the two.

Ridge Regression is a technique to address the problem of multi-co linearity. If we used best subset as a way of dropping the unnecessary model complexity, then we used the *Ridge regression* technique. Both the *lasso* and *ridge regression* are called shrinkage methods. [5]

LASSO technique is applied to perform regularization and variable selection on a predictor model. Depending on the size of the term, LASSO shrinks less applicable predictors to (possibly) zero. Hence, it enable us to consider a more economical model. [6]

Elastic Net technique is applied for an continuation of the lasso that is robust to highest correlations among the predictors. The elastic net uses a mixture of the Lasso and Ridge.

II. REGRESSION TECHNIQUES

In this paper we discuss linear, logistic, stepwise, ridge, lasso, elastic net and implements Linear and Logistic Regression are as follows.

Linear Regression

In Linear Regression technique, we have to establish a relationship model between two variables. One of these variable is called predictor variable whose value is gathered through experiments, these two variables are related through an equation, where exponent (power) of both these variables is 1. The mathematical equation for a linear regression is

[NC-Rase 18]
DOI: 10.5281/zenodo.1483978
ISSN 2348 – 8034
Impact Factor- 5.070

y = ax + b

Algorithm:

- Take out the test of gather a sample of observed values of AverageCoveredCharges, AverageTotalPayments.
- Create a relationship model using the **lm()** functions in R.
- Find out the coefficients.
- Calculate a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- To predict the AverageTotalPayments given to patient, use the **predict()** function in R

Logistic Regression

Logistic Regression is a classification algorithm, is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables, predicts the probability of occurrence of an event by fitting data to a logit().It is component of a larger class of algorithms known as Generalized Linear Model (glm).The fundamental equation of generalized linear model is:

$$g(E(y)) = \alpha + \beta x1 + \gamma x2$$

Algorithm:

- Take out the experiment of gathering a sample of observed values Number.of.Records and AverageCoveredCharges
- Bind a relationship model using the cbind**()** functions in R.
- Here glm() does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.
- The dependent variable need not to be normally distributed.
- Errors need to be independent but not normally distributed.
- Get a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- To predict the charges given to patient, use the **predict()** function in R.

Stepwise Regression:

It includes regression models in which the option of predictive variables is accepted by automatic procedure. Stepwise selection has two main approaches as the forward selection, backward elimination and a combination of the two.

Algorithm:

- Take out the experiment of gathering a sample of observed values of bwt.
- bwt is predefine dataframe
- Here glm() is a modeling function that fits generalized Linear Models.
- Get a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- Every arguments in the stepAIC() function are set to default. If you want to set direction of stepwise regression, the direction argument should be assigned. By default is both.

Ridge Regression:

It is a technique to address the problem of multi-co linearity. If we used best subset is, reducing the unnecessary model complexity, this time we used the *Ridge regression* technique. Here both the *lasso* and *ridge regression* are called shrinkage methods. Shrinkage methods has all predictor values but regularize them towards zero.

Algorithm:

- Take out the experiment of sample of observed values from medicare database.
- glm() is a modeling function that fits generalized Linear Models

[NC-Rase 18] ISSN 2348 – 8034
DOI: 10.5281/zenodo.1483978 Impact Factor- 5.070

- The glmnet() function takes an alpha argument that determines what method is used. If alpha=0 then *ridge regression* is used
- Find the coefficient using the coef() in R
- To predict the contents given to patient, use the **predict()** function in R.

LASSO Regression

Least Absolute Shrinkage and Selection Operator (LASSO) is applied to performs regularization and variable selection on a model. Depending on the size of the term, LASSO shrinks less applicable predictors to (possibly) zero.

Algorithm

- Take out the experiment of sample of observed values from medicare database.
- Produce a relationship model using the **lm()** functions in R.
- Find the coefficients from the model created and create the mathematical equation
- This glmnet() function takes an alpha argument that determines what method is used. Dissimilar values of alpha return different estimators, alpha = 1 is the lasso.
- To predict the contents given to patient, use the **predict()** function in R.
- Calculate the mean
- The MSE is a bit higher for the lasso estimate. Let's check out the coefficients.

Elastic Net Regression

Elastic Net technique is applied for an continuation of the lasso that is robust to highest correlations among the predictors. This technique solves this regularization problem.

Algorithm:

- Take out the experiment of gathering a sample of observed values
- Package to fit ridge/lasso/elastic net models
- Set seed() for reproducibility
- Split data into train and test sets
- Here glmnet() does not assume a linear relationship between dependent and independent variables.
- α is strictly between 0 and 1, and a nonnegative λ, elastic net solves the problem where Elastic net is the same as lasso when $\alpha = 1$. Here α shrinks toward 0, elastic net approaches ridge regression.
- Fit the models
- For plotting , type plot.

III. COMPARISON OF REGRESSION TECHNIQUES

This paper deals with various regression techniques applied on Medicare databases. We studied different techniques of regression which are applied on large database. The results have been produced using R Statistical software.But on this database these two techniques are working properly.

Sr no	Regression Technique	Min	1Q	Median	3Q	Max
1.	Linear Regression	-110702	-10114	-2370	7633	185318

Sr. No	Regression Technique	Std Error	t-value	Residual Std Error	Multiple R-Square	Adjusted R-Square	F-Stat	p-value
1.	Linear Regre	3.561	145.34	20480	0.6787	0.6787	2.112	< 2.2

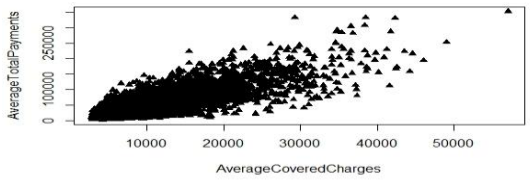
[NC-Rase 18]
DOI: 10.5281/zenodo.1483978
ISSN 2348 – 8034
Impact Factor- 5.070

	ssion							

- The first technique is Linear Regression, it provides Scatter Plot shows strong Linear Trend which is quite High. Hence we fit the Line Of Regression of Y on X.We see the output of Summary function we see that point estimates are highly significant. The Coefficient

Determination labeled as Multiple R-Squared. The F statistics given in the end is a square of t-statistics for testing.

The ROC curve, evaluating the results of classification according to the Predicted values comes from Average Covered Charges and Average Total Payments



The second technique is Logistic Regression, it provides the output that it is significantly associated with the probability of taking Average covered charges and Number of records of the patient. The results obtained with the "R" software show a very good model.

Sr no	Regression Technique	Median	Std Error	z-value	Null Deviance	Residual Deviance	AIC
1.	Logistic Regression	0.0167	8.008	0.990	38.068	35.998	39.998

Residual

Sr no	Regression Technique	Min	1Q	Median	3Q	Max
1.	Logistic Regression	-4.0773	0.0069	0.0167	0.0255	0.0.0448

Classification for the estimation sample

Table (ActualValue=datavar$Number.of.Records,predictedvalue=res>0.3)

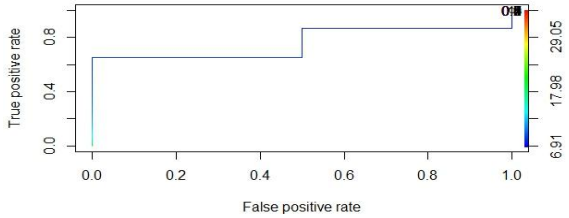
predictedvalue

- ActualValue TRUE

0 2

ISSN 2348 – 8034
Impact Factor- 5.070

The ROC curve, evaluating the results of classification according to the Predictive Accuracies is Number.of.Records and AverageCoveredCharges with its true and false positive rates



IV. COMPARISION

Sr No	Regression Techniques	Standard Error
1	Linear Regression	3.561
2	Logistic Regression	8.008

The regression methods Linear and Logistic having standard error 3.567 and 8.008 on medicare dataset. On this dataset Linear Regression Technique is good than Logistic regression Technique because standard error in Linear Regression is Less than Logistic Regression.

V. CONCLUSION

In this paper, we examined the performance of the six regression techniques used for Big Data Analytics on Medicare dataset. We presented algorithms of these methods through regularized profile plots. The results for the regression technique suggest that we may observe performance differences with these algorithms. We have compared Standard Error for these two techniques and these techniques are used on Medicare dataset

REFERENCES

1. Gemson Andrew Ebenezer J.1 and Durga S.2," BIG DATA ANALYTICS IN HEALTHCARE: A SURVEY" ARPN Journal of Engineering and Applied Sciences ©2006-2015 Asian Research Publishing Network (ARPN). All rights reserved. VOL. 10, NO. 8, MAY 2015, ISSN 1819-6608
2. Manpreet Singh, Vandan Bhatia, Rhythm Bhatia," BIG DATA ANALYTICS Solution to Healthcare" 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT) Manipal University Jaipur, Dec 22-23, 2017, 978-1-5386-3030-3/17/$31.00 ©2017 IEEE
3. M. Zachsenhouse, An Info-Gap Approach to Linear Regression, 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings
4. CHAO-YING JOANNE PENG," An Introduction to Logistic Regression Analysis and Reporting", The Journal of Educational Research • September 2002
5. C. Saunders, A. Gammerman and V. Vovk, Ridge Regression Learning Algorithm in Dual Variables
6. Chris Fraley and Tim Hesterberg, Least-Angle Regression and LASSO for Large Datasets.