IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol-6, Special Issue-6, July 2018 E-ISSN: 2347-2693

Statistical Predictabilty in Big Data Analytics with Data Partitioning

K. Saritha[1*], Sajimon Abraham[2]

[1]School of Computer Sciences, Mahatma Gandhi University, Kottayam, India
[2]School of Management and Business Studies, Mahatma Gandhi University, Kottayam, India

[*]*Corresponding Author: sarithakris@gmail.com, Tel.: +91-94956-86350*

Available online at: www.ijcseonline.org

Abstract— The huge volumes of data which cannot be manipulated easily by commonly available tools are termed as Big Data. Big Data analytics gives competitive opportunities in designing business plans for Business Analytics. The results are used for taking intelligent business decisions; hence it must be accurate and well-timed. For analytical purpose we use Multiple Linear Regression (MLR) model in the statistical method, a type of Supervised Machine Learning Algorithm. Performance of the particular MLR model with one quantitative dependent attribute and four independent attributes are evaluated using splitting up of the whole data set with Cross-Validation technique. This technique is used to validate the accuracy of model developed from training data with test data to control the problem like over fitting. Here we use Hold-Out Cross Validation method with serial and random partitioning. The data set from UCI machine learning repository are evaluated through simulation methods to check the performance. The model generated in training data are validated with test data, the evaluation shows that the result obtained is a generalized one. The proposed MLR model can be used in the new data set for an accurate result. Here we obtained that the accuracy, measuring with random partitioning is a better method.

Keywords— Big Data Analytics, Multiple Linear Regression, Predictive Analytics, Validation Methods

I. INTRODUCTION

The widespread use of digital technologies has led to the exponential growth of data from every imaginable source such as sensors, purchase transactions and social media networks. The large volume of complex and growing data generated from many distinct sources led to the era of Big Data. Companies depend on this massive data to take intelligence decisions as well as to gain a powerful competitive advantage. Modern society is also impacted by big data involving business, management, medical healthcare and government. Big data is extremely valuable to produce productivity in business and evolutionary breakthroughs in scientific disciplines, which give us a lot of opportunities to make great progress in many fields. In large volumes of data so much useful values are hidden and that can be generated only through the careful analysis. For this a new scientific paradigm has been born as Data Intensive Scientific Discovery (DISD) also known as Big Data Analysis.

Big Data can be characterized by the following aspects: (1) volume of data is huge, (2) the data cannot be stored in the regular relational databases, (3) data generation, capturing and processing must be performed very quickly [9]. Big Data analytics technologies and techniques should analyze the huge volume of data and generating conclusion from them to enhance the business and customer relationship [2]. So BI is a technology based analytical process for analyzing huge amount of data and presenting the information to help the end users to increase the decision making capacity. The data used in BI is historical data as well as newly generated data from distinct sources. BI programs are the combination of following advanced analytic techniques: data mining, predictive analytics, statistical analysis, text mining, and Big Data Analytics.

Now a day's the analytical techniques uses data mining, statistics and machine learning for analyzing the huge data set for BI. Indeed the researchers are trying to develop new techniques and technologies or improving the existing one for enhances the BI tools. The aim of the business forecasting tools is to automatically discover the hidden useful information from the large data set.

Rest of the paper is arranged as follows: section II covers related works, section III gives a detailed study on prediction with multiple linear regression, section IV describes the partitioning techniques with linear regression for finding the accuracy of the model, section V contains the implementation methods and the results and its discussion are detailed in section VI. We conclude with a brief discussion in the last section.

II. RELATED WORK

Hai Wang et al. [10] presents in their work the challenges and new trends faced by Big Data while taking decision making. C L Philip Chen and Chu-yang Zhang [9] in their paper has discussed about various methods to manage the flooded data, modular, cloud, bio-inspired and quantum computing. Muhammad Bilal et al. [12] give detailed review on the topic Big Data as Big Data Engineering and Big Data Analytics. Here the authors focus on Statistics, Mining and Artificial Neural Network technologies. Astird Schneider et al. [8] in their review article discuss the performance and interpretation of linear regression analysis. Regression models such as simple multivariable and multivariate are detailed here. Ahmet A. Yildirim et al [7] demonstrate parallel data reduction techniques for Big Data set.

III. PREDICTION WITH MULTIPLE LINEAR REGRESSION

In fact statistics is an essential part of Big Data analytics. With the help of statistical modeling dataset describes the causal relationship among two attributes or along with several attributes[12]. The predictable attribute is called the dependent variable or the response variable. While the attribute that is used for predicting the cause of response variable is called independent variable or predictor variable. Statistical dependences between the attributes give the measure of association among the attributes. Many kinds of Statistical models are used in the different disciples for the purpose of model building and testing, one finds a collection of perceptions about the relationship between causal explanation and observed prediction [1]. The process of predictive model consists of different steps such as Define the goal, Data collection and management, Data preprocessing, Exploratory data analysis, Build the model and Interpreting the result [2].

In the era of Big Data we need efficient statistical methods are needed for predicting the outcomes [3]. Predictive modelling is a commonly used statistical technique to predict future behaviour and is a form of data-mining technology that works by analyzing historical and current data and generating a model to predict future outcomes. Exploratory Data Analysis (EDA) is an important step in the process of statistical predictive analysis. Data exploration uses a combination of summary statistics and the visualization of the data. Summary statistics means find out the mean, variance, median, min, max and quartiles of the quantitative data [4]. From the summary we can find outliers. The outlying data diverges from whole data is known as outliers. This can significantly changes the results of the data analysis and statistical modeling. The use of graphics to examine data is called visualization. Different graphical representations are best suited for answering different questions. In a predictive situation, one might discover the numerical summary of all variables. The correlation among the response variable and other independent variables are examined to find the role of the attributes.

In statistical predictive modelling different kinds of regression techniques are used to predict the result with the help of independent attributes and dependent attributes. Linear Regression model is a supervised machine learning model used for predictive analytics in statistics. This model evaluates the performance based on the correlation and causal relationship among different attributes in the data set. After carryout the steps up to EDA in predictive analytics we can derive the dependent or response variable Y and independent variables from n objects in the data set [5]. The Multiple linear regression (MLR) model is represented with the formula,

Y□□β□□□β□□X□□□β□X□□□□□□□□□□□□β□X□□
□□ε □□□

Where Y is the response variable and X_1, X_2 … X_v are independent variables. β_0, $\beta_{1\dots}$ β_v are unknown regression parameters and ε is called the random error or the noise. The linear regression model must estimate β□ˆ, β□ˆ□□□□□□□□□□□□□β□ˆ□□with□□the given data set and we can predict the future data using the following formula:

Y ˆ
□□β□ˆ□□β□ˆ□X□□□β□ˆ□X□□□□□□□□□□□□□β□ˆ
X□□□□ε̂□□□

To evaluate the performance of a statistical regression model on a new data set we need to find how well its prediction is performed on the existing data set. In regression the coefficient β and Mean Squared Error (MSE) are estimated on the observed data. The estimation of a good regression model is based on the variance between Y ˆ and Y. Also it based on the Bias, is the rate of error in the estimation of a regression model. The rate of change of these two quantities determines whether the test MSE will increase or decrease. So the Bias-Variance trade-off is referred to as the relationship between Bias, Variance, and MSE. We have to build a model with low bias and low variance and then the model should decrease the MSE value.

The regression coefficients $\hat{\beta_i}$, i = 0,1,2,…..v, is estimated and these values minimizes the MSE. With the help of these coefficients the accuracy of the multiple linear regression models is estimated. The accuracy of MLR model is also estimated with two quantities such as Root Mean Squared Error (RMSE) and R^2 statistics [6].

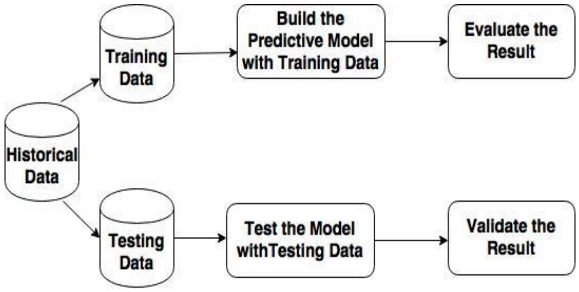
The RMSE value is used to determine the lack of fit of the proposed regression model in the observed data. If the predicted value $\hat{Y_i}$ with the proposed model is very close to the actual dependent variables value Y_i, that is $\hat{Y_i} \cong Y_i$, i=1,2,…..v, then the RMSE value is very low. The RMSE value is very small, and then we can conclude that the proposed model is fit for the collected data very well. The RMSE value is calculated with respect to Y. The formula for calculating the RMSE value is:

RMSE = √ (1/n) ∑ (Y_i -$\hat{Y}$ $_i$) 2□□□□□i =1,2…..n (3)

The R^2 value is an alternate quantity to measure the fitness of the model. The coefficient of determination, R^2, is a measure of how well the regression model describes the observed data. Actually the R^2 statistics is the measure of linear relationship among Y and X. In multiple linear regression models, R^2 is the square of correlation coefficient. If the R^2 statistic value is near to 1, it indicates that a large portion of the collected data is fall in the linear relationship of the proposed model.

IV. PARTITIONING THE BIG DATA SET

Relaying on traditional statistics for Big Data Analytics will not always gives a good result. Hence we propose a new approach that is splitting method in modern statistics with supervised learning method for Big Data Analytics. In this technique we usually split the original data set into training set and testing set similar to machine learning. Here the training data set is used for build the MLR model and estimate variability of the model with different parameters. Then we can test the predictability of the model with the test data set. In traditional statistical analysis build a linear regression model with the sample set of whole dataset. If we build it that way, there is no way to tell how the model will perform with new data. So the preferred practice is to split our dataset into training and test, then build a MLR model on the training sample and then verify the predictability of the dependent variable using test data [8]. The Fig. 1 shows the overview of this model.



.

Fig. 1 Schematic Representation of this model

Cross-Validation is a splitting technique can be often used to estimate the error rate associated with a given statistical learning method and estimating the expected prediction error. One of the Cross-Validation techniques is Hold-Out Cross-Validation. Hold-Out Cross-Validation dividing the available set of observations into two parts, a training set and a validation set or hold-out set. Usually two-third of the data is used for training set and remaining portion is used for test set or Hold-Out set. The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the test set. Fig. 2 shows the representation of this method.

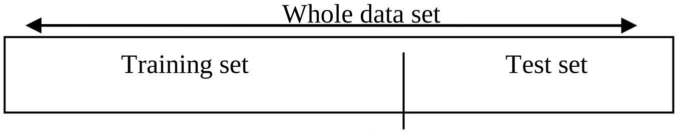


Fig. 2 A schematic display of Hold-Out method

Here we evaluate the performance of the proposed model with the training data using the estimation of the regression coefficients and RMSE value. R^2 value is an alternate measure for determine the fitness of the proposed model. R^2 value tells only the percentage of the information in the dataset or variation in the outcome explained by the model, but not the accuracy. Advantage of the hold-out strategy is that it fully independent of data; only needs to be run once so has lower computational costs. However, its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the test set, and thus the evaluation may be significantly different depending on how the division is made.

V. IMPLEMENTATION

The data set available in UCI machine learning repository is examined here. In the selected data set two wheelers are given to customers on hourly basis. The customers can collect the vehicle from any point and can return the same at any other point. The data set contains information like customer usage details, weather details, monthly usage, holiday usage etc. The size of the data set is 16 into 17379 [14].

Our objective is to build a good statistical model that will predict the number of bikes is needed at any given hour of a particular day. Here we suggest the supervised learning methods for finding the count of bikes at a particular time. For this, at first we serially partition the data set in two: a) two third of the data in the training set and b) the remaining one third as test data. Then the same thing is repeated with partitioning the data set randomly.

Then the steps in the predictive analytics such as pre-processing the data set and EDA are performed on the training data and we find out the response variable Y as “cnt” and the independent variables X as “tem”, “humd”, “Windspeed” and “reg”. The response variable “count” represents the total count of the rental bikes including both casual and registered users. The independent variables “tem”, “humd” and “reg” represents temperature, humidity and the registered users count and “wind” represents the normalized wind speed of a particular hour. The regression model gives the practical relationship among the response variable and the independent variables. This allows the user to decide which of the independent variables have cause on the response. The correlation coefficient is calculated with these independent variables with the response variable “cnt”. From this correlation coefficient we can see that how much each independent attributes is positively or negatively correlated with the response variable [5].

At first we split the whole data set serially into two sets and following steps are performed. Now we build Multiple linear regression model with "tem", "humd", "windspeed" and "reg" are the independent variable and "count' is the response variable with training data. The regression coefficients, RMSE value and R^2 value is estimated on this data. Then check the fitness of the model with the remaining test data that we create at the time of serial partitioning and evaluate the predictability.

The values of regression coefficients are greater than zero in the sets of data, which means that the each independent variable is contributing to the dependent variable "cnt". But in the case of RMSE values there is a huge difference between these two data sets. The RMSE value is used to determine the fitness of the proposed regression model in the observed data. Here we observed that the RMSE value of the test data set is greater than the RMSE value of the training data. This says that our proposed model is badly over fit the data.

The performance of the proposed MLR model is again tested with randomly partitioned data set. Here we get the RMSE value in the two data sets and there is only a small difference between two. So we propose that in supervised learning method the random splitting of the dataset is more accurate. The simulation and results are discussed here.

VI. RESULTS AND DISCUSSION

The results are evaluated and the performance of the proposed work is estimated with different spitting strategies are as follows:

A. Evaluation 1

In the first splitting strategy we split the data set in to two as serially as training data and test data. The training data set contains the two third of the original data set, that is 11586 instances, and the test data contains one third of the whole data set, that is 5793 instances. Now we build a MLR model on the training data set with the following equation:

$$cnt = \beta_0+\beta_1 temp+ \beta_2 hum+ \beta_3 wind+ \beta_4 registered \quad (4)$$

Here all regression coefficients values are greater than zero. So we found that there is a relationship between the independent variables "temp", "hum", "wind" and "registered" with the response variable "cnt". Table 1 gives the details of coefficients of the regression model between the temperature, humidity, wind speed and registered user with "cnt" in training data.

Table 1. Coefficients of the model

$count = \beta_0+\beta_1 tem+ \beta_2 humd+ \beta_3 wind + \beta_4 reg$

	Estimated value on training data	**Estimated value on test data**
Intercept	6.821	40.81
Tem	75.351	93.79
Hum	-42.101	6.7
Wind	4.771	21.17
Reg	1.117	1.008

The quality of a linear regression is also explained using two related quantities: RMSE value and the R^2 statistics. Here RMSE=33.25 and is relatively small, representing that the model fit the data well. R^2 measures the percentage of predictability in "cnt" that can be explain using the independent attributes. An R^2 statistic that is close to 1 indicates that a high percentage of the variability in the response has been explained by the regression. Here the R^2 =0.949, is close to 1 indicates that a large proportion of the variability in the response has been explained by this regression model. Table 3 gives the values of RMSE and R^2.

The same estimation method is repeated on the testing data and the values of regression coefficients are estimated with same MLR model. Table 1 gives the values of regression coefficients. From these values we found that there are huge differences between the values of each regression coefficients, but all values are greater than zero. So we can conclude that there is relationships between these variables are strong which we found in the model with training data set.

In this model evaluation the RMSE value is 45.2 and R^2 value is 0.9576. The RMSE value is greater than the value we get with test data. But the R^2 value is very close to the values we get from the model with the training data set. That is this particular MLR model will not have similar prediction accuracy with the training data. Table 3 gives the values of RMSE and R^2 values with test data. From this the serial partitioning of the whole data will not give exact prediction accuracy and the proposed model is badly over fit the data.

B. Evaluation 2

In the second evaluation method we partition the data set randomly. That is two third of the data is selected randomly and used for training and remaining one third is also selected randomly and used for test data. At first we build the same MLR model as in the first evaluation with randomly partitioned training data. Here also we get the regression coefficients, which are greater than zero. Table 2 gives the values of the regression coefficients of the above proposed model between the temperature, humidity windspeed and registered user with "cnt" in training data.

Table 2. Coefficients of the model

$cnt = \beta_0+\beta_1 temp+ \beta_2 hum+ \beta_3 wind + \beta_4 registered$

	Estimated value on training data	Estimated value on test data
Intercept	14.958	11.355
Tem	85.33	84.049
Hum	-60.788	-56.382
Wind	1.635	0.687
Reg	1.104	1.115

The quality of this MLR model is also evaluated with RMSE value and R^2 statistics. With this data splitting we get RMSE= 38.12 and R^2=0.9565. Table 3 gives details of RMSE value and R^2 values with training data.

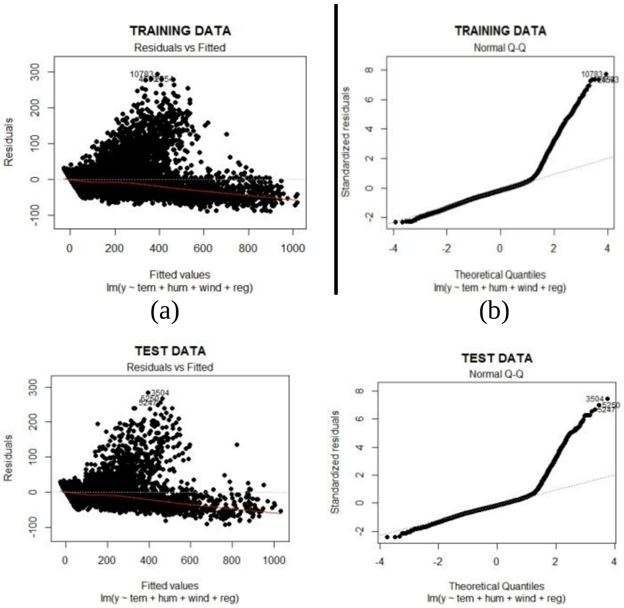
The same methods are repeated with the testing data. Now we get the regression coefficients, which are similar values with the training data. Table 2 gives the estimated values of regression coefficients with test data. Here we found that there is very negligible difference between each regression coefficients with the training data and testing data. So we can adopt this splitting method as a supervised learning method for MLR model and we suggest that this MLR model have good predictability.

The RMSE value and R^2 statistics are also estimated with the training data and we get RMSE=38.2 and R^2 = 0.9543. These two values have very slight difference from the corresponding values we get with training data. The details are given in Table 3.From Table 3 we can compare each of RMSE value and R2 values. In the case of serial partitioning the difference between RMSE value between the training data and test data is high compared the same with the random partitioning. So we propose that in supervised learning method random partitioning of the data set is a better method and the MLR model is created with training data and the accuracy is tested by over-fitting the model with the test data.

Table 3. Details of RMSE and R^2 values

Data	RMSE	R^2
Training data with serial partition	33.25	0.949
Test data with serial partition	45.2	0.9576
Training data with random partition	38.12	0.9565
Test data with random partition	38.2	0.9543

Another method for checking the accuracy of the proposed MLR model is the residual analysis. Residuals are used to represents how efficiently the proposed model using the data. Different kinds of residual plots are used for calculate and judge the proposed MLR model. Fig.3 shows different types of residual plots with the training data and test data. Fig. 3 (a) and (c) are the Residuals vs Fitted graph is used to find the randomly distributed points which are fit for the training data and test data respectively.



(a) (b)

(c) (d)

Fig. 3 Visualization plots for MLR model with training data and test data

Fig.3 (b) and (d) are Normal Q-Q graph is a probability graph. From these graph we can found that majority of the points in the graph is fall in the straight line. So the normality test is satisfied with the training data. It is cleared from the Fig. 3 (b). It can be tested with the test data and we get a similar plot in Fig. 3(d). Here we implement supervised learning technique with MLR model and we found that the proposed model is fitted in the data set.

VII. CONCLUSION and Future Scope

Different splitting strategies of better analysis of big data are suggested in this paper. Here data set is divided serially and randomly as training and test set and the accuracy of the particular MLR model is evaluvated. The model is created with the taining data and the accuracy is tested by overfitting the model with test data. The results obtained are then verified and it seen that in random partitioning the difference between the values of regression coefficients, RMSE value and R^2 statistics obtained from the training data and test data are very negligible. So we conclude that validation test with supervised learning method and random partitioning will give an accurate result.

REFERENCES

[1] Kumar, P., & Rathore, D. V. S. (2014). "Efficient capabilities of processing of big data using hadoop map reduce". International Journal of Advanced Research in Computer and Communication Engineering, 3(6), 7123-6..

[2] Feldman, D., Schmidt, M., & Sohler, C. (2013, January). "Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering". In Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 1434-1453). Society for Industrial and Applied Mathematics.

[3] Ha, S., Lee, S., & Lee, K. (2014). "Standardization Requirements Analysis on Big Data in Public Sector based on Potential Business Models". International Journal of Software Engineering and Its Applications, 8(11), 165-172.

[4] Galit Shmueil, "To Explin or Predict?", Statistical science, vol25 © Institute of Mathematical Science, 2010

[5] Saritha, K., & Abraham, S. (2017, July). "Prediction with partitioning: Big data analytics using regression techniques". In Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on (pp. 208-214). IEEE.

[6] Dutta, P. S., & Tahbilder, H. (2014). "Prediction of rainfall using data mining technique over Assam". Indian Journal of Computer Science and Engineering (IJCSE), 5(2), 85-90.

[7] Ahmet A Yildirim, Cem OZdogan, Dan Watson, " Parallel Data Reduction Techniques for Big Data sets", Research gate, 2016.

[8] Astrid Scheneider, Gerhard Hommel and Maria Blettner, "Linear Regression Analysis", 2010; 107(44) 776-82

[9] Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314-347

[10] Wang, H., Xu, Z., Fujita, H., & Liu, S. (2016). "Towards felicitous decision making: An overview on challenges and trends of Big Data". Information Sciences, 367, 747-765.

[11] Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., ... & Pasha, M. (2016). "Big Data in the construction industry: A review of present status, opportunities, and future trends", Advanced Engineering Informatics, 30(3), 500-521.

[12] Saritha, K., & Abraham, S. "Big Data Challenges and Issues: Review on Analytic Techniques". Indian Journal of Computer Science and Engineering (IJCSE) Vol. 8 No. 3 Jun-Jul 2017

[13] https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Authors Profile

Mr Sajimon Abraham. MCA, MSc(Mathematics), MBA, PhD (Computer Science). He has been working as Faculty Member in Computer Applications & IT, School of Management and Business Studies, Mahatma Gandhi University, Kottayam, Kerala, India and also research guide in Mahatma Gandhi University. He currently holds the additional charge of Director (Hon), University Center for International Co-operation. He was previously working as Systems Analyst in Institute of Human Resource Development, Faculty member of Computer Applications in Marian College, Kuttikkanam, Kerala, India and Database Architect in Royal University of Bhutan under Colombo Plan on deputation through Ministry of External Affairs, Govt of India. His research area includes Data Science, Spatio-Temporal Databases, Mobility Mining, Sentiment Analysis, Big Data Analytics and E-learning and has published 52 articles in National, International Journals and Conference Proceedings.

Ms Saritha K pursed Bachelor of Science from Mahatma Gandhi Universty, Kottayam, Kerala, Master of Computer Application from Bharathidasan University, Trichy, Thamilnadu and M.Phil (Computer Science) from Bharathiyar University, Tamilnadu. She is currently pursuing Ph.D. and working as Faculty Member in Department of Computer Science, School of Technology and Applied Sciences, Kottayam, Kerala. She has 17 years of teaching experience and two years of research experience. Her main research area focuses on Data Science, Data Mining Statistical Data Analytics and Neural Networks.