2015 Fifth International Conference on Communication Systems and Network Technologies

Correlation Analysis of Big Data to Support Machine Learning

Dr. Rajiv Pandey
Amity Institute of Information Technology
Amity University
Lucknow , Uttar Pradesh, India
rajivpandeylko@gmail.com

Manoj Dhoundiyal
IT Department
Amity University
Lucknow , Uttar Pradesh, India
mdhaundiyal@amity.edu

Amrendra Kumar
IT Department
Amity University
Lucknow , Uttar Pradesh, India
akumar9@amity.edu

***Abstract*— The large size and complexity of datasets in Big Data need specialized statistical tools for analysis and we use R for correlation analysis of our data set. This paper explores the correlation analysis through best fit linear regression of quantitative variables with help of the demonstration based on scatter plots and linear regression best fit line. The analysis demonstrated in this paper is scalable to Big Data in any other context where the quantitative variables are clearly delineated. R provides multiple techniques and inferences to statistical analysis of dataset, this paper however explores the correlation between quantitative variable establishing the extent of dependability between them using R functions. The correlation and best fit line functions of R i.e. cor( ) and abline(lmout) respectively are significantly explored.**

Keywords—Quantitative Variables; R; Correlation analysis; Big Data; Linear Model, Linear Regression

I. INTRODUCTION

Big data is related to that dataset which exhibits attributes like dynamicity, unstructured data, and also posses categorical and quantitative components. Big data comes from various resources ranging from web, social media, click stream data, sensors and other connected devices [1] enhancing the volume. The social networking applications, e-commerce and Business intelligence applications are major sources of data generation.

The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions [2].

The data analysis of such a huge datasets is a complex challenge, the multiple quantitative variables, velocity, volume and veracity of the dataset adds to the complexity. The data size is large and knowledge extraction challenges can be overcome by various data analytic tools like R, Octave, Hive and associated techniques.

R is open source programming language and software environment designed for statistical computing and visualization [3].

This paper explores the analytical inference that can be drawn out of the dataset by establishing relationship between the various quantitative variables using R and its function of correlation and linear regression. The paper is divided into five sections Section 1 being the Introduction, Section 2 : "Big Data: features and dimensions" helps us to list the aspects which differentiate it from any other dataset/databases, this section also list the sources of data contributing to Big Data in Education, Section 3: "Features of Quantitative Variables", helps us to understand which tool to be deployed and what summary features can be used to draw inferences on such a huge data set, Section 4 of the paper "Correlation and its significance in Machine Learning" describes Correlation and the various statistical measures. Section 5 "Statistical Inference using R" demonstrates the features of the data set under consideration by closely analyzing the relationship using R.

II. BIG DATA: FEATURES AND DIMENSIONS

A. Big data characteristics

Big data and its analysis are of prime importance when the dynamism of the dataset multiplies million times within hours. The analysis or generation of this size was not thought anticipated of and the importance of analyzing such trend was not dealt in the past.

Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and centralized control and seeks to explore complex and evolving relationships among data [4], HACE theorem.

Big Data is generally characterized by the three V's i.e. Volume, Variety, Velocity, but of late the Big Data has demonstrated certain additional dimensions.

The Features of Big Data [5] can be summarized as:

- Volume: size growing many folds with each minute passing by.
- Variety: Multi format Data.
- Velocity: The dynamics at which it is multiplying.
- Veracity: Confirming to facts and the fear in the minds of decision makers.
- Complexity: Multi device point of generation. Their integration is a challenge.

Big Data is also identified in the C3 Space [6] so that the Storage, Mining, Data analytics and Machine Learning aspects can be mathematically modeled and statistically analyzed for better utilization. The three C are

- Cardinality: Defines the number of records.
- Continuity: including the characteristics of data representation and data size.
- Complexity : includes three dimensions as
  - Large variety of data types.

978-1-4799-1797-6/15 $31.00 © 2015 IEEE
DOI 10.1109/CSNT.2015.32

- High dimensional dataset.
- Demand of high speed data processing.

The size and complexities involved with Big Data is a major issue but if effective data analytical tools deployed it can also be advantageous in business, decision making and predictions.

B. Big data in education

The education has transformed from offline/Indoors to online. MOOC's is a burning example of this trend where students across the world are educated online in a virtual classroom. The student data size has multiplied manifolds to the tune of millions. The education has changed from chalk and talk to mobiles and PDA's. The requirements of the online students have changed w.r.t time, use and availability of technology, and scope of the subject. It therefore is of significant concern to address the needs, queries and provide tailor made curriculum and suggestion to pursue desired and suitable courses.

Big Data sources [5] in education are:

- Documents in non-electronic form,
- Data from information systems,
- Logs from university servers,
- Opinions from social networks
- Data from public education portals.

The sources being numerous the input and size of data becomes really Big, therefore all the aspects of Big Data, Machine Learning and predictions shall apply to education domain like it may apply to any other area. The size and dimension being so diverse it calls for a tool that can handle the dataset and generate valuable outcomes.

The data analytics in education is suggested using a Big Data analytical tool R in the subsequent sections of the paper, primarily focusing on the correlation of quantitative variables.

III. Features of Quantitative Variables

Categorical variables and quantitative variables both play a significant role in the analysis of any data set, there are multiple statistical measures in respect to quantitative variables but only a few exist when it comes to categorical variables. Quantitative variables can be converted in to categorical variables as marks listed below are categorized as:

30-50	50-60	60-70	80 and above
Poor	Average	Good	Excellent

"One approach is to represent the categories with numerical values (quantification) prior to visualization using methods for numerical data" [7].

This section lists the features of each to bring the subject in context, however the paper only analyses the relationship of the quantitative variable.

A. Quantitative Variables

Quantitative variables provide a larger set of statistical measures and detail inferences as under

- Minimum and Maximum value
- Quartiles
- Mean, Median and Mode
- Standard Deviation
- Box Plot and scatter plot analysis.
- Histogram representation
- Extreme values
- Skew measures or Symmetric/non Symmetric behavior

The following section deliberates the correlation and linear regression on a dataset of education domain.

IV. Correlation and its significance in Machine Learning

After Correlation is a statistical technique that demonstrates how strong is the bonding of a pair of variables. It works for quantifiable data. Correlation in conjunction with linear regression a supervised machine learning (ML) technique.

If x and Y are two quantitative variables having values from 1 to n and

If yi = xi or yi = a.xi + b (where a>0)
then correlation is positive tending to +1.
else if yi = a.xi + b (where a<0)
then correlation is negative tending to -1.
else correlation is 0
Which means two variables are not related.

$$\text{Correlation} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Fig 1. Expression to evaluate correlation between x and y

Correlation has a mathematical formula Fig. 1 which when substituted with the values of the variables generates correlation. In this paper we have demonstrated the correlation findings through R tool and we have not indulged in mathematical solutions.

Correlation can have a value:

- 0 > correlation <=1 (i.e. positive value), which means it is a perfect positive correlation and means that both variables tend to increase together.
- 0, signifies no correlation meaning the variables x and y have no relation what so ever.
- -1 <= correlation < 0 (i.e. negative value), which means it is a perfect negative correlation and means that when one variable increase the other one decreases.

Correlation like the line of best fit is useful in capturing the linear relationship of the quantitative variables.

In correlation each observation helps us to determine which of the two variables is effecting the other one and to what extent. The variable that impacts is called as exploratory variable and the one which is affected is called the response variable. Based on extent of impact and relationship and we can plan an appropriate analysis.

Scatter plot is a very common graphic visualisation tool which helps us to study the relationship and infer whether the correlation exists or not, if it exists whether it is positive or negative. This relationship and further analysis is described in the subsequent sections.

V. STATISTICAL INFERENCE USING R

For a comparative study of two quantitative variables, we analyzed the dataset of approximately 300 students. This dataset contains the marks obtained by these students in their Matriculation exam as one of its quantitative field and marks obtained in Aptitude test during `their higher studies as the second field apart from other columns. These two marks are then analyzed for finding correlation between them with the help of R-tool. Scatter plot is drawn with Matriculation marks at X-axis and Aptitude marks at Y-axis.

>plot(data[,2],data[,3],xlab="Marks obtained in Matriculation", ylab="Marks in Higher Education (Aptitude Test)")

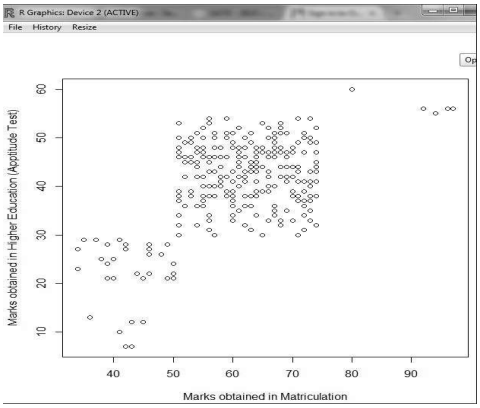

Fig 2. Scatter Plot

In section IV (Fig 1) we have listed the mathematical derivation of correlation whereby a detailed solution is required to generate the correlation coefficient. The R tool however performs this task using cor( ) function. cor ( ) function (Fig 3) of R programming is used to calculate the correlation between quantitative variables.

>cor(data[,2],data[,3]) cor() function is supplied with two arguments i.e. data[,2] and data[,3], which fetch the columns two and three i.e. Matriculation Marks and Aptitude Marks respectively from our dataset. The variable data has been initialized with the dataset using read.csv() function of R.

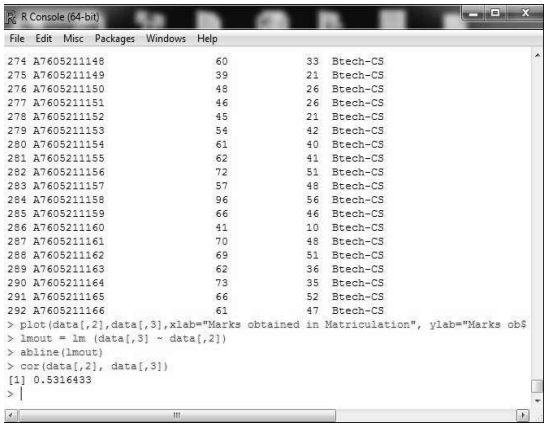

Fig 3. Output of cor() function in R

The positive output of cor( ) function which is 0.5316, establishes the fact that both variables are in positive correlation and tend to increase together.

The Linear regression best fit line can be generated using R that shall help to infer the positive or negative relationship between target and explanatory variable. This line can support ML applications to gather business and domain intelligence for decision making. The regression line is generated in the subsequent section using the above dataset.

The above graphic visualization can be further enhanced to infer more meaningful analysis. This is done by incorporating a best fit line demonstrating the extent of correlation. The lm() functionality of 'R' is used to perform regression analysis. The lm( ) function (short for "Linear Model") is also capable to calculate the residuals, hypothesis test statistics and many more values, other than the estimated coefficients. Output of lm( ) function [8] is added as line of best fit in our scatter graph with the help of abline( ) function. The intercept and the slope of the abline() are influenced by the features of the quantitative variables which are passed as vector argument, lmout, calculated using lm() function of R.

>lmout = lm(data[,3] ~ data[,2])

>abline(lmout)

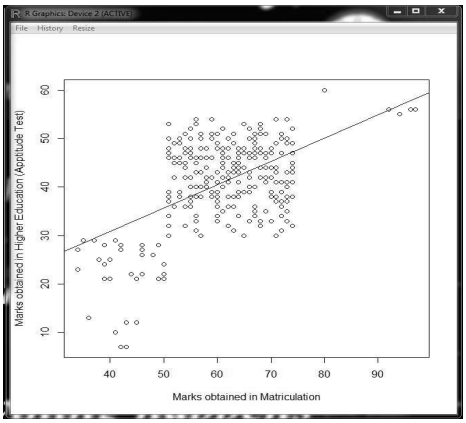

Fig 4. Scatter plot along with line of best fit

The regression best fit line is based on the considered dataset and stores all of the variables, correlation coefficient and other features that may be significant for predicting and identifying data patterns to support the target variables. In our context the best fit line would help the machine to predict the relationship between matriculation marks and marks obtained in aptitude test i.e. the explanatory and target variables respectively. Since the coefficient is positive, and the regression line having a positive slope, it is clearly inferred that higher the marks in tenth, higher are the chances to obtain higher marks in aptitude test.

Fig. 5 summarizes the model parameters of the dataset stored in the variable data which are the resultant of summary(lmout) [9] command of R. The residual standard error of 7.737 is considered to be on the higher side meaning that the chances of prediction based on best fit line is likely to produce greater number of false alarm, which is not a desirable input for machine learning.

>summary(lmout)

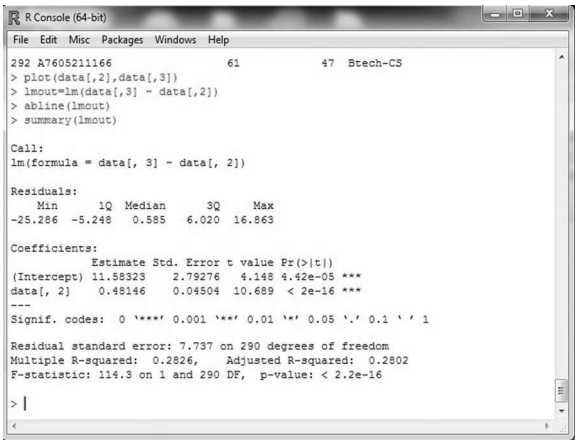

Fig 5. Model Parameters

The analytical inferences resulting from the above description can be summarized as

- Higher the number of marks in matriculation exam, higher are the marks scored by the candidate in Aptitude test.
- Correlation [Fig 3] shows a positive trend between matriculation marks and Aptitude test of Higher education.Since the intercept of line is positive and line rows towards North-East direction, we can conclude that quantitative variables are in positive relation.
- The correlation analysis has been performed using various R functions such as cor(), lm() and abline().
- The linear regression summary measures generated using summary(lmout) is helpful in providing model parameters like degree of freedom , residual standard error etc., which significantly impact statistical inferences.

The above summarized parameters can significantly help in predictive Analytics to determine futuristic relationship between the variables.

Predictive Analytics [11] is used to:

- Predict future trends and probabilities
- Analyze relationships in data not visible with conventional analysis

The above statistical correlation analysis can be used as a major input for machine learning algorithm which can subsequently make predictions and inferences as the need be.

VI. CONCLUSION

The paper has explored the comparative study of quantitative variables using a contemporary data analytical environment R and the correlational aspect has been visualized using the functions of R. The best fit line, the resultant of the function abline(lmout) can be used to make predictions based on correlation i.e. positive or negative correlation. The statistical inferences described, can be used to train a Machine for predictions and classification through adapting various machine learning algorithms.

REFERENCES

[1] G. Mark Kerzner and Sujee Maniyam, “Hadoop illuminated” https://github.com/hadoop-illuminated/hadoop-book

[2] A. Rajaraman and J. Ullman, “Mining of massive data sets”, Cambridge Univ. Press, 2011.

[3] Wei Fan and Albert Bifet, “Mining big data: current status and forecast to the future” SIGKDD Explorations ,Volume 14 issue 2.

[4] Xindong wu, Xingquan Zhu,Gong-Qing and Wel Ding, “Data Mining with big data”, IEEE Transactions on knowledge and data engineering,Vol 26,No.1,January 2014.

[5] Peter Michalik, Jan Stofa and Iveta Zolotova, “Concept definition for big data architecture in the education system” IEEE 12th international Synposium on Applied Machine Intelligence and Informatics,January 2014.

[6] Shan Suthaharan, “Big data classification:problems and challenges in network intrusion prediction with machine learning”, Performace Evaluation Review,Vol. 41, No 4 March 2014 .

[7] Sara Johansson Fernstad and Jimmy Johansson “A task based performance evaluation of visualization approaches for categorical data analysis” IEEE DOI 10.1109/IV.2011.92, 2011

[8] Norman Matloff “The Art of R Programming – A tour of statistical software design”, 2011.

[9] Vignesh Prajapati “Big data analytics with R and Hadoop” PACKT Publishing Ltd. ISBN 978-1-78216-328-2 Nov 2013

[10] W. N. Venablles, D.M. Smith and the R Core Team, “An introduction to R” , Notes on R: A Programming Environment for Data Analysis and Graphics,2013.

[11] G.C Deka, “Handbook of research on cloud infrastructures for big data analytics”,IGI Global,370-391,20