PREDICTIVE ANALYTICS USING STATISTICS AND BIG DATA:
CONCEPTS AND MODELING

Krishna Kumar Mohbey
Arvind Pandey
Dharmendra Singh Rajput

Bentham Books

Predictive Analytics Using Statistics and Big Data: Concepts and Modeling

Edited by

Krishna Kumar Mohbey
Central University of Rajasthan
India

Arvind Pandey
Department of Statistics
Central University of Rajasthan
India

&

Dharmendra Singh Rajput
VIT Vellore
India

Predictive Analytics Using Statistics and Big Data: Concepts and Modeling

Editors: Krishna Kumar Mohbey, Arvind Pandey and Dharmendra Singh Rajput

ISBN (Online): 978-981-14-9049-1

ISBN (Print): 978-981-14-9051-4

ISBN (Paperback): 978-981-14-9050-7

need for a court order if at any point you breach any terms of this License Agreement. In no event will any delay or failure by Bentham Science Publishers in enforcing your compliance with this License Agreement constitute a waiver of any of its rights.

3. You acknowledge that you have read this License Agreement, and agree to be bound by its terms and conditions. To the extent that any other terms and conditions presented on any website of Bentham Science Publishers conflict with, or are inconsistent with, the terms and conditions set out in this License Agreement, you acknowledge that the terms and conditions set out in this License Agreement shall prevail.

Bentham Science Publishers Pte. Ltd.
80 Robinson Road #02-00
Singapore 068898
Singapore
Email: subscriptions@benthamscience.net

CONTENTS

FOREWORD

Big data analytics started receiving increasing attention a few years ago. It all began by exploring how to deal with the increasing volume, variety, and velocity of the data. Today, developing effective and efficient approaches, algorithms, and frameworks are used as they are essential to deal with such data.

Earlier, data were limited, and experiments were also performed for limited scopes. With the advent of big data, internet technologies, massive information, and predictive analytics problems have exploded in complexity and behavior. There are increasing challenges in the area of data storage, management, and computations. There is a need to combine various researches related to big data technologies, statistics, and predictive analytics into a single volume.

The proposed eBook addresses a comprehensive range of advanced topics in big data technologies with statistical modeling towards predictive analytics. This book will be of significant benefit to the community as a useful guide of the latest research in this emerging field, *i.e*., predictive analytics. This ebook will help the studies in this field finding relevant information in one place.

R.S. Thakur
Maulana Azad National Institute of Technology
Bhopal
India

PREFACE

Predictive analytics is the art and science of proposed predictive systems and models. With tuning over time, these models can predict an outcome with a far higher statistical probability than mere guesswork. Predictive analytics plays an essential role in the digital era. Most of the business strategies and planning depend on prediction and analytics using statistical approaches. With the increasing digitization day by day, analytical challenges are also increasing at the same rate—digital information, which is rapidly growing, generating vast amounts of data. Hence, the design of computing, storage infrastructures, and algorithms needed to handle these "big data" problems. Big Data is collecting and analyzing complex data in terms of volume, variety, and velocity. The most extensive selection of big data is from digital information, social media, IoT, sensor, *etc*.

Predictive analytics can be done with the help of various big data technologies and statistical approaches. Big data technologies include Hadoop, Hive, HBase, and Spark. There are numerous statistical approaches to perform predictive analytics, including Bayesian analysis, Sequential analysis, Statistical prediction, risk prediction, and decision analytics.

This book presents some latest and representative developments in predictive analytics using big data technologies. It focuses on some critical aspects of big data and machine learning and provides descriptions for these technologies.

The book consists of seven chapters. Chapter 1 discusses data analytics in multiple fields with machine learning algorithms. An application of bootstrap sampling is presented in chapter 2 with the case study of quantifying player's monopoly in a cricket team. Successive sampling for mean estimation is discussed in chapter 3. Chapter 4 discussed log type estimators of population mean under ranked set sampling. Bivariate survival data analysis is represented in chapter 5. An approach for weblog data analysis using machine learning techniques is discussed in chapter 6. Chapter 7 discussed an epidemic analysis of COVID-19 using exploratory data analysis approaches.

Many eminent colleagues made a significant impact on the development of this eBook. First, we would like to thank all the authors for their exceptional contributions to the eBook and their patience for the long process of editing this

eBook. We would also like to thank the reviewers for their insightful and valuable feedback and comments that improved the book's overall quality.

Krishna Kumar Mohbey
Central University of Rajasthan
India

Arvind Pandey
Central University of Rajasthan
India

&

Dharmendra Singh Rajput
VIT Vellore
India

List of Contributors

Anoop Kumar	Department of Mathematics and Statistics, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India
Arvind Pandey	Department of Statistics, Central University of Rajasthan, India
Bireshwar Bhattacharjee	Department of Statistics, Assam University, Silchar, Assam, India
Brijesh Bakariya	Department of Computer Science and Engineering, I. K Gujral Punjab Technical University, Punjab, India
Chemmalar Selvi G.	School of Information Technology and Engineering, VIT University, Vellore, India
Dharmendra Singh Rajput	VIT Vellore, India
Dibyojyoti Bhattacharjee	Department of Statistics, Assam University, Silchar, Assam, India
Krishna Kumar Mohbey	Central University of Rajasthan, India
Lalpawimawha	Department of Statistics, Pachhunga University College, Mizoram, India
Lakshmi Priya G.G.	School of Information Technology and Engineering, VIT University, Vellore, India
Nishi Rastogi	Department of Statistics, National P. G. College, Lucknow, India
R. Suguna	Sri Sarada College for Women (Autonomous), Salem-16, Tamilnadu, India
R. Uma Rani	Sri Sarada College for Women (Autonomous), Salem-16, Tamilnadu, India
Shailja Pandey	Department of Mathematics and Statistics, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India
Shashi Bhushan	Department of Mathematics and Statistics, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India
Shikhar Tyagi	Department of Statistics, Central University of Rajasthan, India

CHAPTER 1

Data Analytics on Various Domains with Categorized Machine Learning Algorithms

R. Suguna*, R. Uma Rani

Sri Sarada College for Women (Autonomous), Salem-16, Tamilnadu, India

Abstract: Data Analytics is an emerging area for analyzing various kinds of data. Predictive analytics is one of the essential techniques under data analytics, which is used to predict the data gainfully with machine learning algorithms. There are various types of machine learning algorithms available coming under the umbrella of supervised and unsupervised methods, which give suitable and better performance on data along with various analytics methods. Regression is a useful and familiar statistical method to analyze the data fruitfully. Analysis of medical data is most helpful to both patients as well as the experts to identify and rectify the problems to overcome future problems. Autism is a brain nerve disorder that is increasing in the children by birth due to some most chemical food items and some side effects of other treatments and various causes. Logistic Regression is one of the supervised machine learning algorithms which can operate the dataset of binary data that is 0 and 1.

Agriculture is one of the primary data which should be considered and analyzed for saving the future generation. Rainfall is a more elementary requirement for the global level and also countries which are having backbone as agriculture. Due to the topography, geography, political, and other socio-economic factors, agriculture is affected. Thus, the demand for food and food products is intensifying. Especially crop production is depending upon the rainfall, so, prediction of rainfall and crop production is essential. Analysis of social crime relevant data is indispensable because analytics can produce better results, which leads to reducing the crime level. Unexpectedly child abuse is increasing day by day in India. Linear regression is the supervised machine learning algorithm to predict quantitative data efficiently.

This chapter is roofed with various datasets such as autism from medical, rainfall, and crop production from agriculture and child abuse data from the social domain. Predictive analytics is one of the analytical models which predict the data for the future era. Supervised machine learning algorithms such as linear and logistic regression will be used to perform the prediction.

*Corresponding author R. Suguna: Sri Sarada College for Women (Autonomous), Salem-16, Tamilnadu, India; Tel: +91-4274550291; E-mail: sugunarmca@gmail.com

Krishna Kumar Mohbey, Arvind Pandey & Dharmendra Singh Rajput (Eds.)

Keywords: Data analytics, Exponential distribution, Inomial distribution, Linear regression, Logistic regression, Machine learning, Normal distribution, Prediction analytics.

INTRODUCTION

Data Analytics

Analysis of data is a need and essential task to get the solutions for the problems. In our society, there is plenty of progress going, and data are increasing day by day. Analytics helps to analyze the data. There are various real kinds of domains, such as medical business, agriculture, and crimes, which are the most considerable areas. Analytics is advanced mining of data with the standard statistical and mathematical methods. There are various kinds of analytical methods available, they are,

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

Descriptive and diagnostic analytics concentrated on past events, which analyze the past data. Predictive and prescriptive analytics are used to give future solutions for past data [1].

Machine Learning

Fig. (**1**) shows the classification of machine learning algorithms. Machine learning is apart from the root of artificial intelligence. A machine can read and analyze the data. There are some wings of machine learning that are supervised, unsupervised, and reinforcement learning.

Supervised Machine Learning

Classification

Classification is the process of classifying the data with the pre-specified rules. The classifier's labels are predetermined. Prediction can be made using the classification and regression techniques. There are various famous machine learning algorithms available for classification. They are listed below.

Random Forest

- It is an ensemble method, and it reduces the over-fitting of the result. This algorithm is mainly used for prediction. From the training data, voting is used to get the prediction.

Decision Trees

- Decision trees are another effective method for classification. It is a tree-like model consisting of root and leaf nodes. It can handle high dimensional data and provides accuracy at a reasonable level.

Nearest Neighbour

- It is a simple classification algorithm that classifies the nearest samples based on the likeness of the data points. Choosing the k value is most important in this algorithm. Based on the cross-validation or else square root of the samples, k value will be better on the training data.

Boosted Trees

- Boosting is a process of making a compilation of adapting weak learners to powerful learners. It is an ensemble method, and the weak learner is a classifier that is scarcely correlated with classification, and a powerful learner is a classifier with a strong correlation.

Support Vector Machine

- SVM is a robust classification algorithm that classifies the data on the N-dimensional space, and N is the number of features. SVM will find a hyperplane to classify the data. The hyperplane is a boundary for the classification of data.

Neural Networks

- The neural network is working based on the biological neuron. The layers of neural networks are the input layer, an output layer, and the hidden layer. The input layer takes the samples of input data, and processing of data is done by the hidden layer, then the calculated output is produced by the output layer.

CHAPTER 2

Quantifying Players' Monopoly in a Cricket Team: An Application of Bootstrap Sampling

Bireshwar Bhattacharjee* and **Dibyojyoti Bhattacharjee**

Department of Statistics, Assam University, Silchar, Pin: 788011, Assam, India

Abstract: Cricket is a bat-and-ball game. It is played between two teams, each team consisting of 11 players. In limited-overs cricket, the teams play for a fixed number of overs, usually 50 or 20. At the end of the match, the team which scores the most number of runs in those limited-overs win the match. In this paper, taking the data from ICC Cricket World Cup 2019, an attempt is made to identify the type of competition that exists between the players, *i.e.*, batsmen and bowlers using the Herfindahl-Hirschman Index (HHI). This index is a statistical device used for estimating the degree of concentration in a particular market. A team is said to have the monopoly in scoring runs if the bulk of their scoring in the tournament is done by a few batsmen only while the other batsmen made an insignificant contribution with the bat. Likewise, a team is said to have bowler's monopoly, if the majority of wickets is taken by few bowlers of the team while the other bowlers could dismiss an insignificant number of opponent batsmen in the tournament. Applying bootstrap sampling, the teams are classified into three groups *viz*. monopoly, moderately competitive, and perfectly competitive. From the analysis, it is found that India, Australia, Bangladesh, and New Zealand are the teams where a monopoly exists, *i.e.*, most numbers of runs are scored by two or three batsmen. All other teams except Pakistan, *i.e.*, Afghanistan, West Indies, Sri Lanka, South Africa, England, are categorized as having perfect competition in the task of run-scoring. On the other hand, in the case of bowlers, Australia, Pakistan, Sri Lanka, and Bangladesh enjoys monopolistic nature in bowling. All other teams such as India, New Zealand, West Indies, Afghanistan, South Africa, and England are categorized as having Perfect Competition in the task of taking wickets. The study finds that out of the four semi-finalists, three of the teams enjoy a monopoly of batsmen, and three teams enjoy the perfect competition of bowlers. Thus, the work concludes that the monopoly of batsmen in a cricket team and perfect competition amongst bowlers have a role to play in the performance of teams in the tournament.

Keywords: Bootstrapping, Cricket, Cricket analytics, HHI, ICC cricket world cup, monopoly.

*__Corresponding author Bireshwar Bhattacharjee:__ Research Scholar, Department of Statistics, Assam University, Silchar, Pin-788011; Tel: +91-9531045037; Email: bireshwarbhattacharjee09@gmail.com.

Krishna Kumar Mohbey, Arvind Pandey & Dharmendra Singh Rajput (Eds.)

INTRODUCTION

Cricket is a bat-and-ball game. The game is played in three different formats known as Test, ODI, and Twenty20 Cricket. Test cricket is the longest format, which runs to five days and two innings each team. Twenty20 cricket is the shortest form of cricket, with a maximum of twenty overs each side. It is played between two teams, each team consisting of 11 players. One-team bats (Team A, say) while the other team fields and bowls (Team B, say). Team A tries to score as many runs as possible against the bowling of Team B, who tries to dismiss the batsmen and hence minimizes the runs scored by Team A [1].

One Day International (ODI) is a form of limited-overs cricket, played between two teams with international status, in which each team faces a fixed number of overs, usually 50. The first ODI was played on 5th January 1971 between Australia and England at the Melbourne Cricket Ground. The One Day Cricket World Cup is generally held every four years. The first World Cup was organized in England in June 1975, which was won by West Indies. Australia lifted the trophy for a maximum number of five times, West Indies, and India lifted the trophy two times, and Pakistan, Sri Lanka, and England won it once [1].

In this paper, an attempt is made to identify the major contributory batsmen and bowler and their impact on their team's performance. This study is based on data from the ICC Cricket World Cup 2019 because after 1992, the type of format was first introduced where each team plays against every team once, and there would be 9 games for each team in the group stage since the participating teams are restricted to 10.

This paper is divided into seven subsections. The first subsection deals with the introduction; this is followed by the motivation of the study. Third section deals with the review of the literature. The objectives of the study are illustrated in the fourth subsection. Fifth subsection deals with methodology. Results and discussion are provided in the sixth subsection. The last section elaborates on the conclusions and directions for further research.

MOTIVATION OF THE STUDY

Cricket is a team game, where batting and bowling are the prime skills of the game. A better team shall have several good batsmen as well as superlative bowlers of different specialization who perform as and when demanded to make sure that their team wins against their opponent. But having few quality batsmen and bowlers in

the team may not always reflect in the team performance as other team members may not perform efficiently, leading to excessive dependence of the team on few players. This may lead to a situation where the chance of victory of a team gets reduced. Thus, it has been felt to design a study to ascertain the type of competition that exists between the teams that participated in the Cricket World Cup of 2019 in terms of their batting and bowling performance.

REVIEW OF LITERATURE

In this section, we try to discuss some literature related to the quantification of the performance of batsmen, bowlers, and that of teams' performance. Though in cricket, there are several available measures to quantify the performance of players, like batting average and strike rate for batsmen, bowling average, bowling strike rate and economy rate for bowlers and percent of victory for teams; such measures always do not reveal the true level of performance of the cricketers and teams. Accordingly, different researchers have contributed significantly to defining several measures of performance analysis in cricket.

Kimber and Hansford [2] observed that the batting average while quantifying the batsman performance is ubiquitous in cricket, so they concluded that the traditional batting average depends on an unrealistic parametric assumption. They proposed a non-parametric approach based on runs scored for assessing batting performance. Some other works related to batting performance are conducted by Wood [3], Barr and Kantor [4] Maini, and Narayanan [5] Borooah and Mangan [6], *etc*.

Lemmer [7], while analyzing the performance of players, developed a current bowling performance measure (CBR), which is a joint measure that takes into account all the important measures of bowling performance. Its calculation requires the use of an easily programmable algorithm using only the bowler's career values of overs bowled, runs given, and wickets per innings played. Some other works related to bowling performance were performed Beaudoin and Swartz [8], Garber and Sharp [9], Dey *et al*., [10] Bhattacharjee *et al*., [11], *etc*.

Passi and Pandey [12] attempted to predict the performance of players- batsman and bowlers. This is done by classifying the number of runs scored, and the number of wickets dismissed in different ranges. They have used naive Bayes, random forest, multiclass SVM and decision tree classifiers to generate the prediction models for both the categories.

CHAPTER 3

On Mean Estimation Using a Generalized Class of Chain Type Estimator under Successive Sampling

Shashi Bhushan[1], Nishi Rastogi[2] and **Shailja Pandey[1,*]**

[1] *Department of Mathematics and Statistics, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India*
[2] *Department of Statistics, National P.G. College, Lucknow, India*

Abstract: The present paper comprises a significantly generalized class of chain type estimators to estimate the population means on the current occasion under the framework of successive sampling based on auxiliary information on both the occasion. The proposed generalized class constitutes two renowned chain type classes proposed by Singh and Vishwakarma [1, 2]. As its particular case, an improvement over their notion, with some eased regularity conditions, is proposed by us, which consists of chain type regression estimators additionally to chain type ratio estimators. The construction of the proposed class is fruitful in the sense of constructing the chain type classes of estimators in the realm of successive sampling. In terms of efficiency, we provide a comparative study of the proposed class oversample mean estimator, Cochran's estimator [3], Sukhatme *et al.* estimator [4] and Singh's estimator [5]. A numerical illustration is demonstrated in support of the proposed class.

Keywords: Auxiliary information, Generalized class of chain type estimator, Optimum replacement policy, Successive sampling.

INTRODUCTION

In practice, the structure of the population under study remains stable over the given period while in applied sciences, sociology, and economic researches, this composition of the population changes over the same span. In such situations, successive sampling over different points of time equips a useful statistical tool for estimating the reliable numerical estimates. A common phenomenon to improve the precision on the most recent occasion in successive sampling, we use the estimates obtained on earlier occasions. This approach of drawing sample on the successive occasion was set up by Jessen [6] in the analysis of an agricultural survey

[*]**Corresponding author Shailja Pandey:** Department of Mathematics and Statistics, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India; Tel: +91-7271020995; E-mail: writetoshailja6693@gmail.com.

Krishna Kumar Mohbey, Arvind Pandey & Dharmendra Singh Rajput (Eds.)

data. For estimating more precise estimates on the current occasion, he availed all the information available. A similar notion of estimation was lengthened by Patterson [7], Rao, and Graham [8], Sen [9], Gupta [10], Das [11], Chaturvedi and Tripathi [12], and among others. It has been observed that the auxiliary information in sample surveys plays a supporting key role in enhancing the efficiency of the estimators. Therefore, besides information available on the previous occasion, continuous sampling yields more efficient estimates in the existence of auxiliary information. In many circumstances, for the survey purpose, this information may be available readily or made available by adjusting the small part of the available cost. The use of such type of information on two occasions successive sampling for estimating the population mean has been made by Feng and Zou [13], Biradar and Singh [14], Singh and Singh [15], Singh [16], Singh [5] and Singh and Vishwakarma [1, 2].

An earnest effort has been made in the sense of chain aspect, which covered both chain type ratio and regression estimators, in presenting a more general approach than Singh and Vishwakarma [2]. The presented approach also enhances the idea of constructing a chain type class of estimators in successive sampling. Now, Singh and Vishwakarma's [1, 2] classes are the subclasses of the proposed generalized class in *Subsection 2.2*. The coverage of numerous estimators under the proposed generalized chain class is given in *Subsections 2.3* and *2.4*. The minimum mean square error (MSE) of the proposed class is discussed in *Section 3*. Analytical and empirical studies are included in *Sections 4* and *5,* respectively.

SAMPLING METHODOLOGY

Sample Structure and Notations

Let us consider a framework of population size N , which we have to sample over two successive occasions with x and y as the characteristics understudy on the first and second occasions, respectively. Further, we have considered z the auxiliary information readily obtainable on the first and second occasions. For simplicity, we provide results for a large population with a sample of size n on both occasions. By adopting simple random sampling without replacement on the first occasion, we draw a sample of size n . Further, on the second occasion, we retain a subsample of size $m = n(1-\theta)$ from the sampled units of the first occasion and $u = n\theta$ units from the units that were not selected on the first occasion, *i.e.*, $N-n$ units. We shall use the following notations for further use:

$\bar{X}$ and $\bar{Y}$ **:** The population means of study variable x and y on the first and second occasions, respectively.

$\bar{Z}$ **:** The population mean of the readily available auxiliary variable on the first and second occasions both.

$\bar{y}_u$ and $\bar{y}_m$: The sample means of the study variable on the second occasion based on sample sizes u and m respectively.

$\bar{x}_m$ and $\bar{x}_n$ **:** The sample means of the study variable on the first occasion based on sample sizes shown in subscripts m and n , respectively.

ρ_{yx} , ρ_{yz} and ρ_{xz} : These are the correlation coefficients between the variables written in their subscript.

S_y^2 , S_x^2 and S_z^2 : These are sampling variance of the variables displayed in subscript.

θ **:** The unmatched proportion of the sample units

FORMULATION OF THE PROPOSED GENERALIZED CLASS

Following Singh and Vishwakarma [1, 2] and Bhushan and Rastogi [17], we formulate a new generalized class of chain type estimator to estimate the population mean on the current occasion as follows:

$$T = \phi T'_H + (1-\phi)T'_G \qquad \textbf{(1)}$$

here, T'_H is the class of estimators for unmatched proportion as considered by Singh and Vishwakarma [1]

$$T'_H = H'(\bar{y}_u, \bar{z}_u) \qquad \textbf{(2)}$$

Where H' is such that $H'(\bar{Y}, \bar{Z}) = \bar{Y}$ also satisfies the following conditions in a bounded, closed convex subset with the real space $\Re$ of dimension two:

(i) H' the function is continuous and bounded in $\Re$.

CHAPTER 4

Log Type Estimators of Population Mean Under Ranked Set Sampling

Shashi Bhushan and **Anoop Kumar***

Department of Mathematics and Statistics, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India

Abstract: This paper considers some log type and regression cum log type class of estimators under ranked set sampling. The suggested class of estimators are found to be better than most of the estimators proposed to date and equally efficient to the usual regression estimator under ranked set sampling. The theoretical findings have been furnished with a simulation study carried out over some artificially generated symmetric and asymmetric populations. Also, following McIntyre [1], Dell [2], and Dell and Clutter [3], we have investigated the effect of skewness and kurtosis over the efficiency of the proposed class of estimators.

Keywords: Bias, Efficiency, Kurtosis, Mean square error, Ranked set sampling, Skewness.

INTRODUCTION

McIntyre [1] mooted a cost-effective method in the theory of survey sampling and referred to it as a method of unbiased selective sampling using a ranked set, which consolidated the ease of non-probability sampling *via* curb of simple random sampling (SRS). It works as a stratification of samples rather than the classical approach, which preludes the stratification of the population. He showed theoretically that the mean of quantified units gives an unbiased estimator of a population mean irrespective of error in the ranking of units. However, the much expected stringent mathematical foundation to the theory of McIntyre's method was furnished by Takahasi and Wakimoto [4] and referred to it as a method of unbiased estimation of population mean on the sample stratified by means of ordering. The method of McIntyre [1] became inert nearly fourteen years until Halls and Dell [5] administered a field test for the estimation of weights of browse and herbage in a

*Corresponding author Anoop Kumar: Department of Mathematics and Statistics, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, U.P., India; Tel: +91- 8009554263; E-mail: anoop.asy@gmail.com.

Krishna Kumar Mohbey, Arvind Pandey & Dharmendra Singh Rajput (Eds.)

pine-hardwood forest. The first named it as ranked set sampling (RSS) and seen through the empirical observation that it was dominating the SRS. Dell and Clutter [3] pointed out that the RSS estimator of a population mean, is still unbiased with an equal sample size regardless of error involved during the judgement order. Muttlak and McDonald [6] developed RSS for the case when units are selected with size biased probability with respect to the concomitant variable. Muttlak and McDonald [7] evoked an efficient line intercept procedure under RSS. Muttlak [8] considered RSS for the estimation of parameters in simple linear regression. Samawi and Muttlak [9] showed, with the help of real data, which ranking of which variable increases the estimator's efficiency. Kadilar *et al*. [10] suggested a ratio estimator to estimate population mean under RSS. Al-Hadhrami [11] by following Kadilar and Cingi [12], suggested ratio type estimators of population mean under RSS. Al-Omari *et al*. [13] considered the new ratio estimators of population mean using SRS and RSS. Jeelani and Bouza [14] investigated a new ratio estimator using the linear combination of median and quartile deviation of the auxiliary variable under RSS, whereas Jeelani *et al*. [15] suggested ratio estimators of population mean utilizing the linear combination of deciles and coefficient of skewness of the auxiliary variable. Saini and Kumar [16] looked into the modified ratio estimator under SRS and RSS. Mehta and Mandowara [17] introduced modified ratio-cum-product estimators under RSS. Jeelani *et al*. [18] proposed a new ratio estimator under RSS based on deciles of an auxiliary variable. Khan *et al*. [19] considered a new regression cum ratio type estimator under RSS. Recently, Bhushan and Kumar [20] investigated some optimal classes of estimators under RSS. In this paper, we have adopted log type and regression cum log type class of estimators under RSS, employed by Bhushan *et al*. [21] and Bhushan and Gupta [22-26] under SRS. We have observed that in order to find out the expression of MSE, Al-Hadhrami [11] and Khan *et al*. [19] utilized the optimum value of optimizing scalar $\beta = \rho xy(Sy/Sx)$, an optimum value of b under SRS which is irrelevant for use in RSS. So, we have felt to derive the correct MSE by using the optimum value of b under RSS and comparing them with the proposed estimators class.

Further, section 2 is devoted to the methodology of ranked set sampling and review of some existing estimators under *RSS* in the literature. In section 3, the proposed estimators are given along with their properties. In section 4, we have derived the correct expression of *MSE* of Al-Hadhrami's [11] estimator and Khan *et al*.'s [19] estimator. In section 5, we have derived the theoretical conditions under which the proposed class of estimators dominates the other existing estimators. In section 6, a simulation study is added for an illustration. In section 7, the conclusion is made regarding the applicability of the adapted class of estimators. It is noticed that

asymmetry shows an adverse effect on the efficiency of the proposed class of estimators. To verify this fact, we have computed the simulation results under the consideration of different asymmetric distributions.

LITERATURE REVIEW

The concept of ranked set sampling was given by McIntyre [1], which is based on picking out *m* simple random samples each of size *m* units from the population and m units are ranked within each set according to the variable of interest visually or by any cost independent measure. The unit now with *rank 1* is taken for the measurement of the element from the first sample, and the remaining units of the sample are discarded. Again, the unit with *rank 2* is taken for the measurement of the element from the second sample, and the remaining units of the sample are discarded. The process continues in the same way until the unit with *rank m* is taken for the measurement of the element from the m^{th} sample. This above process is defined as a cycle. If this procedure recurred *r* times, then this yield a ranked set sample of size *n=mr* units. Let (X_1, Y_1), (X_2, Y_2),...,(X_n, Y_n) be a bivariate random sample of size *n* with probability density function *f (x, y)* and cumulative distribution function *F(x, y)* having population mean $\bar{X}$, $\bar{Y}$, variances σ_x^2, σ_y^2 and coefficient of correlation ρ_{xy}. Let the ranking be executed on the auxiliary variable x to estimate the population mean of the variable of interest *y*. Let (X_{11}, Y_{11}), (X_{12}, Y_{12}),..., (X_{1n}, Y_{1n}); (X_{21}, Y_{21}), (X_{22}, Y_{22}),..., (X_{2n}, Y_{2n});...;(X_{n1}, Y_{n1}), (X_{n2}, Y_{n2}),..., (X_{nn}, Y_{nn}) be a bivariate random sample drawn from the population with the help of simple random sampling without replacement *(SRSWOR)* having the same *c.d.f. F(x,y)* and $(X_{i(1)}, Y_{i[1]})$, $(X_{i(2)}, Y_{i[2]})$,..., $(X_{i(n)}, Y_{i[n]})$ be the order statistic of X_{i1},X_{i2},..., X_{in} and the judgment order of Y_{i1},Y_{i2},..., Y_{in}; *i=1,2,...,n*. Let $X_{1(1)}$,$X_{2(2)}$,..., $X_{n(n)}$ refers to the ranked set samples.

Samawi and Muttlak [9] suggested the estimator of population ratio as

$$\widehat{\boldsymbol{R}}_{\boldsymbol{r}} = \frac{\bar{\boldsymbol{y}}_{[\boldsymbol{n}]}}{\bar{\boldsymbol{x}}_{(\boldsymbol{n})}} \qquad \textbf{(1)}$$

Samawi and Muttlak [9] mentioned that the above estimator could also be utilized for the estimation of population mean as well as population total. Thus, the estimator for the population mean is

$$\bar{\boldsymbol{y}}_{\boldsymbol{r}} = \frac{\bar{\boldsymbol{y}}_{[\boldsymbol{n}]}}{\bar{\boldsymbol{x}}_{(\boldsymbol{n})}} \bar{\boldsymbol{X}} \qquad \textbf{(2)}$$

CHAPTER 5

Analysis of Bivariate Survival Data using Shared Inverse Gaussian Frailty Models: A Bayesian Approach

Arvind Pandey[1], Shashi Bhushan[2], Lalpawimawha[3,*] and **Shikhar Tyagi[1]**

[1]*Department of Statistics, Central University of Rajasthan, India*
[2]*Department of Mathematics and Statistics, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India*
[3]*Department of Statistics, Pachhunga University College, Mizoram, India*

Abstract: Frailty models are used in the survival analysis to account for the unobserved heterogeneity in individual risks of disease and death. The shared frailty models have been suggested to analyze the bivariate data on related survival times (*e.g*., matched pairs experiments, twin or family data). This paper introduces the shared Inverse Gaussian (IG) frailty model with baseline distribution as Weibull exponential, Lomax, and Logistic exponential. We introduce the Bayesian estimation procedure using Markov Chain Monte Carlo (MCMC) technique to estimate the parameters involved in these models. We present a simulation study to compare the actual values of the parameters with the estimated values. Also, we apply these models to a real-life bivariate survival data set of McGilchrist and Aisbett [1] related to the kidney infection data, and a better model is suggested for the data.

Keywords: Bayesian model comparison, Inverse gaussian frailty, Lomax distribution, Logistic exponential distribution, MCMC, Shared frailty, Weibull exponential distribution.

INTRODUCTION

The statistical analysis of time-to-event, event-history, or duration data plays an essential role in medicine, epidemiology, biology, demography, engineering, actuarial science, and other fields. In the past several years, medical research concerning the addition of random effects to the survival model has substantially increased. The random effect model is a model with a continuous random variable presenting excess risk or frailty for individuals or families. Sometimes due to the

[*]**Corresponding author Lalpawimawha:** Department of Statistics, Pachhunga University College, Mizoram, India; Tel: +91-9862307640; E-mail: raltelalpawimawha08@gmail.com.

Krishna Kumar Mohbey, Arvind Pandey & Dharmendra Singh Rajput (Eds.)

economic reasons or human ignorance or non-availability of the factors, the crucial factors are unobserved in the model. This unobserved factor is usually termed as heterogeneity or frailty. In the statistical modeling concept, the frailty approach covers the heterogeneity caused by unmeasured covariates like genetic factors or environmental factors. In the frailty model, the random effect (frailty) has a multiplicative effect on the baseline hazard function, extending the Cox proportional hazard model. Clayton [2] introduced a random effect model to account for the frailty shared by all individuals in a group.

Generally, this heterogeneity is often referred to as variability in survival analysis and is considered one of the critical sources of variability in medical and biological applications. It may not be easy to assess, but it is nevertheless of great importance in the model. In the model, if the frailty is either statistically impactful or ignored, the model will be unsuitable, and the decision based on such models will be misleading; ignorance of the frailty may lead to either underestimation or overestimation of the parameters and higher values of AIC, BIC, DIC can be seen in comparison to the model (after including frailty) [3]. In general terms, we let the heterogeneity go into the error term. It leads to an increase in the response's variability compared to the case when frailty is included. The unobservable risks are random variables, which follow some distribution. It is possible to choose a different distribution for unobserved covariates. The variance of the frailty distribution determines the degree of heterogeneity in the study population. This paper considers the shared frailty model with a random effect or frailty in the hazard model, common and shared by all individuals in the group [3, 4]. Since the frailty is not observed, we assume to follow a positive stable distribution.

In practice, the gamma frailty specification may not fit well [5-7]. However, gamma frailty has drawbacks. For example, it may weaken the effect of covariates studied by Hougaard [8] in the analysis of multivariate survival data. IG distribution can be another practical choice. Concerning time, the population becomes homogeneous under IG, whereas the relative heterogeneity remains constant for gamma [9], udder quarter infection data; Duchateau and Janssen [10] fit the IG frailty model with the Weibull hazard rate.

The gamma model has predictive hazard ratios that are time-invariant and may not be suitable for these patterns of failures [11].

The term "frailty" itself was first introduced by Vaupel *et al*. [4] in univariate survival models and was substantially promoted by its applications to the

multivariate survival data. In shared frailty models, we assumed that the survival times are conditionally independent for the given shared frailty. That means dependence between survival times is only due to frailty. In the frailty model, the unobserved random effect acts multiplicatively on baseline hazard function, which is assumed to follow one of the parametric distributions like gamma, IG, positive stable, log-normal, and power variance function. Let T be a continuous lifetime random variable, and random variable Z be a frailty variable. The conditional hazard function for given frailty at $t > 0$ is given by

$$h(t|z)=zh_0(t)e^{X\beta} \quad \textbf{(1.1)}$$

where $h_0(t)$ is a baseline hazard function at time $t > 0$. X is a row vector of covariates, and β is a column vector of regression coefficients. The conditional survival function for given frailty at time $t > 0$ is,

$$S(t|z)=e^{-\int_0^t h_0(x|z)dx} \quad \textbf{(1.2)}$$

$$=e^{-zH_0(t)e^{X\beta}} \quad \textbf{(1.3)}$$

Where $H_0(t)$ is the cumulative baseline hazard function at time $t > 0$. Integrating over the range of frailty variable Z having density $f(z)$, we get the marginal survival function as,

$$S(t)=\int_0^\infty S(t|z)f(z)dz$$

$$=\int_0^\infty e^{-zH_0(t)e^{X\beta}}f(z)dz \quad \textbf{(1.4)}$$

$$=L_Z(H_0(t)e^{X\beta}) \quad \textbf{(1.5)}$$

where L_Z (.) is the Laplace transform of the distribution of Z. Once we get the survival function at time $t > 0$ of lifetime random variable for an individual, we can obtain probability structure and the inference based on it.

The remaining article is organized as follows. In Sec. 2, the introduction of the generally shared frailty model is provided; also, we have discussed IG shared frailty models, respectively. In Sec. 3 and 4, we have introduced a shared frailty model and baseline distributions. Different proposed shared frailty models are given in section 5. An outline of model fitting, using the Bayesian approach, is presented in Sec. 6.

CHAPTER 6

An Efficient Approach for Weblog Analysis using Machine Learning Techniques

Brijesh Bakariya*

Department of Computer Science and Engineering, I. K Gujral Punjab Technical University, Punjab, India

Abstract: Information on the internet is rapidly growing day by day. Some of the information may be related to the person or not. The amount of data on the internet is very vast, and it is tough to store and manage. So the organization of massive amounts of data has also produced a problem in data accessing. The rapid expansion of the web has provided an excellent opportunity to analyze web access logs. Data mining techniques were applied for extracting relevant information from a massive collection of data, but now it is a traditional technique. The web data is either unstructured or semi-structured. So there is not any direct method in data mining for it. Here Python programming language and Machine Learning (ML) approach is used from handling such types of data. In this paper, we are analyzing weblog data through python. This approach is useful for time and space point of view because because python has many libraries for data analysis.

Keywords: Data mining, Machine learning, Weblog, Python, World Wide Web

INTRODUCTION

Millions of people use the internet at schools, colleges, offices, houses, and many other places. Moreover, the internet is a standard medium to exchange knowledge from different platforms—web users, *i.e*., that user who is using the web [1]. Web-users just browse the web and get the information with a single click. But accessing desired information is the most challenging task. There is lots of hidden information present in a single weblog. Moreover, the weblog record contains various kinds of information like IP address, URL, Referrer, Time, *etc*. (Fig. **1**) shows a sample format of the weblog [2]. There are various types of file formats for storing weblog, such as World Wide Web Consortium, Internet Information Services, Apache HTTP Server, *etc*. World Wide Web Consortium is a log of web server that contains a text file with different types of attributes, such as IP address, timestamp, the HTTP

*Corresponding author Brijesh Bakariya: Department of Computer Science and Engineering, I.K Gujral Punjab Technical University, Punjab, India; Tel: +91-9465884876; E-mail: brijeshmanit@gmail.com.

Krishna Kumar Mohbey, Arvind Pandey & Dharmendra Singh Rajput (Eds.)

version, the browser type, the referrer page, *etc*. Moreover, Internet Information Services is also a web server from Microsoft that also contains information about logs with various attributes. There are various types of files including different log file format, but analyzing log and getting information from that log is a very challenging task. The Apache HTTP Server clearly provides log information.

```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 200 6245
unicomp6.unicomp.net - - [01/Jul/1995:00:00:06 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
burger.letters.com - - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/countdown/liftoff.html HTTP/1.0" 304 0
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 304 0
burger.letters.com - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/video/livevideo.gif HTTP/1.0" 200 0
205.212.115.106 - - [01/Jul/1995:00:00:12 -0400] "GET /shuttle/countdown/countdown.html HTTP/1.0" 200 3985
d104.aa.net - - [01/Jul/1995:00:00:13 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
129.94.144.152 - - [01/Jul/1995:00:00:13 -0400] "GET / HTTP/1.0" 200 7074
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
unicomp6.unicomp.net - - [01/Jul/1995:00:00:14 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
d104.aa.net - - [01/Jul/1995:00:00:15 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204
129.94.144.152 - - [01/Jul/1995:00:00:17 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 304 0
199.120.110.21 - - [01/Jul/1995:00:00:17 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713
ppptky391.asahi-net.or.jp - - [01/Jul/1995:00:00:18 -0400] "GET /facts/about_ksc.html HTTP/1.0" 200 3977
205.189.154.54 - - [01/Jul/1995:00:00:24 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
ppp-mia-30.shadow.net - - [01/Jul/1995:00:00:27 -0400] "GET / HTTP/1.0" 200 7074
205.189.154.54 - - [01/Jul/1995:00:00:29 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310
ppp-mia-30.shadow.net - - [01/Jul/1995:00:00:35 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 200 5866
205.189.154.54 - - [01/Jul/1995:00:00:40 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
ix-orl2-01.ix.netcom.com - - [01/Jul/1995:00:00:41 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985
ppp-mia-30.shadow.net - - [01/Jul/1995:00:00:41 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786
ppp-mia-30.shadow.net - - [01/Jul/1995:00:00:41 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 200 363
```

Fig. (1). Sample of the weblog.

Machine Learning (ML) is concerned with computer programs that automatically improve their performance through experience. In machine learning, there is a learning algorithm, then data called as training data set is fed to the learning algorithm. The learning algorithm draws inferences from the training data set. It generates a model, which is a function that maps input to the output. Fig. (**2**) shows the process of machine learning. There are various applications of ML, such as Text Categorization, Fraudulent Transactions, Face Recognition, Recommendations, Robot Navigation, Market Segmentation, and many more. There are some important points where learning is used, such as human expertise, does not exist, for example, navigating on mars. Learning is used when humans are not able to explain their expertise; for example, speech recognition [3].

Machine Learning is used when solution changes in time, for example, routing in a computer network. Learning is used when a solution needs to adapt to particular

cases; for example, person biometric; there are many areas where learning can be used [4].

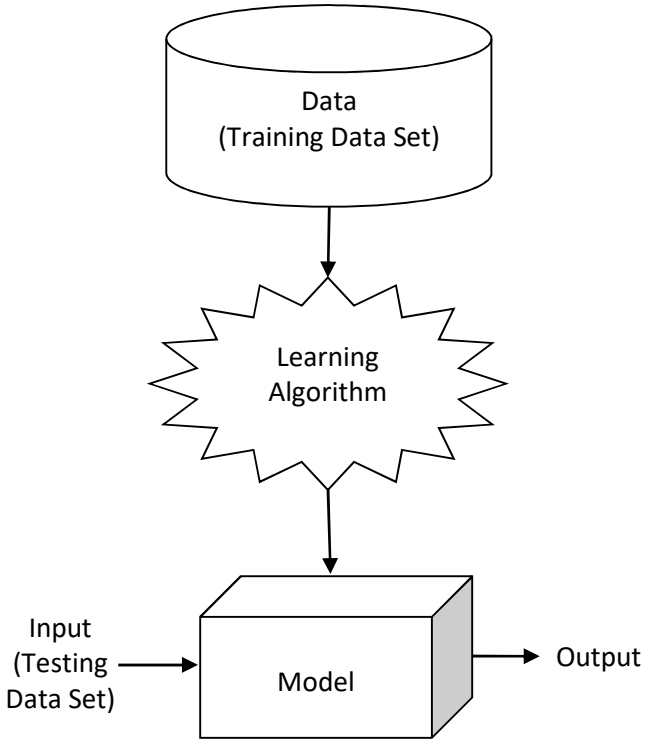

Fig. (2). Process of ML.

MACHINE LEARNING TECHNIQUES

ML algorithms are very efficient for data analysis. In the case of machine learning-based weblog analysis, proposed systems are trained in such using supervised or unsupervised learning algorithms to classify the weblog taken from web servers or other repositories. These are the following types of learning methods:

Supervised Learning

In supervised learning, datasets are divided into two categories, *i.e*., training datasets and test datasets. In the case of weblog analysis, the training datasets are labeled with IP address, User Name, Timestamp, URL, *etc*. After that, training data sets are used to train the system, and then test data sets are used to test the output of the proposed system.

CHAPTER 7

An Epidemic Analysis of COVID-19 using Exploratory Data Analysis Approach

Chemmalar Selvi G. and **Lakshmi Priya G. G.***

School of Information Technology and Engineering, VIT University, Vellore, India

"Data is the new science. Big Data holds the answers." – By Pat Gelsinger

Abstract: The outbreak of data has empowered the growth of the business by adding business values from the available digital information in recent days. Data is elicited from a diverse source of information systems to bring out certain kinds of meaningful inferences, which serve closer in promoting the business values. The approach used in studying such vital data characteristics and analyzing the data thoroughly is the Exploratory Data Analysis (EDA), which is the most critical and important phase of data analysis. The main objective of the EDA process is to uncover the hidden facts of massive data and discover the meaningful patterns of information which impact the business value. At this vantage point, the EDA can be generalized into two methods, namely graphical and non-graphical EDA's. The graphical EDA is the quick and powerful technique that visualizes the data summary in a graphical or pictorial representation. The graphical visualization of the data displays the correlation and distribution of data before even attempting the statistical techniques over it. On the other hand, the non-graphical EDA presents the statistical evaluation of data while pursuing its' key characteristics and statistical summary. Based on the nature of attributes, the above two methods are further divided as Univariate, Bivariate, and Multivariate EDA processes. The univariate EDA shows the statistical summary of an individual attribute in the raw dataset. Whereas, the bivariate EDA demonstrates the correlation or interdependencies between actual and target attributes; the multivariate EDA is performed to identify the interactions among more than two attributes. Hence, the EDA techniques are used to clean, preprocess, and visualize the data to draw the conclusions required to solve the business problems. Thus, in this chapter, a comprehensive synopsis of different tools and techniques can be applied with a suitable programming framework during the initial phase of the EDA process. As an illustration, to make it easier and understandable, the aforementioned EDA techniques are explained with appropriate theoretical concepts along with a suitable case study.

***Corresponding author Lakshmi Priya G.G.:** VIT School of Design, VIT Vellore, India; Tel: +91-9486322772; E-mail: lakshmipriya.gg@vit.ac.in

Krishna Kumar Mohbey, Arvind Pandey & Dharmendra Singh Rajput (Eds.)

Keywords: Bivariate analysis, Data visualization, Exploratory data analysis (EDA), Multivariate analysis, Statistical methods, Univariate analysis.

INTRODUCTION

Is EDA a Critical Task?

At the outset, the phrase “Data Science” is understood to deal only with statistical data modeling and advanced machine learning techniques. But, the venture of data science stems from the essential keystone, which is frequently underrated or obliterated - Exploratory Data Analysis (EDA). The term EDA was coined by John W. Tukey in 1977 [1]. In the abstract view, EDA is the process of analyzing the main characteristics of the dataset and visualizing the summary of the context of data using statistical methods [2]. It is a critical task before dealing with statistical or machine learning modeling since it perceives the background knowledge required to come about a suitable model to solve the problem at hand and arrive at possible results.

In the early years, computer scientists extracted the knowledge or hidden information from the data by using the technique called Knowledge Discovery Process (KDP). (Fig. **1**)Step 2 of the KDP technique [3] shows data preprocessing where the actual data is cleaned and transformed to make it in a more consistent and understandable format that can be used for inferring knowledge after applying the data mining algorithm.

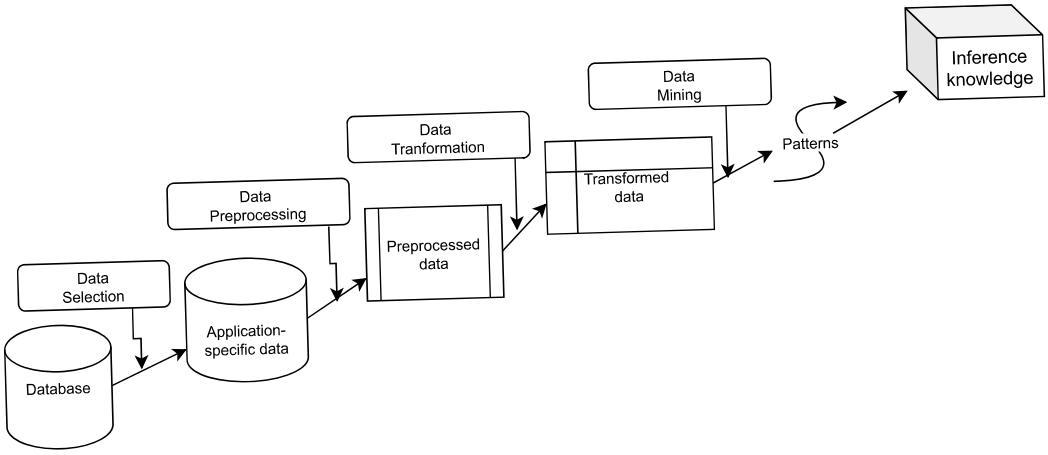

Fig. (1). KDP technique [4].

In recent years, the data scientist unearthed the unknown facts through the EDA process, which is the most important phase of the data science process (Fig. **2**). EDA is highly helpful to the data scientist because of the solutions that they turned out are logically correct and closer to solve the business problems. Apart from turning out with logical solutions, EDA also addresses the business stakeholders by certifying that the right questions are interrogated without ignoring the assumptions and problem statements so as to maximize the benefit of the data scientist's result. This EDA process is a perk to the business stakeholder or data scientist since it provides the more significant intuition behind the data that would not even be thought of to investigate, yet it can be highly demanded insights to the business problems. Hence, EDA is a process or technique which is used to examine the dataset to assess the patterns, identify the relationship, describe the data characteristics, and visualize the statistical summary about the data. It is applied before any data-driven model is constructed.

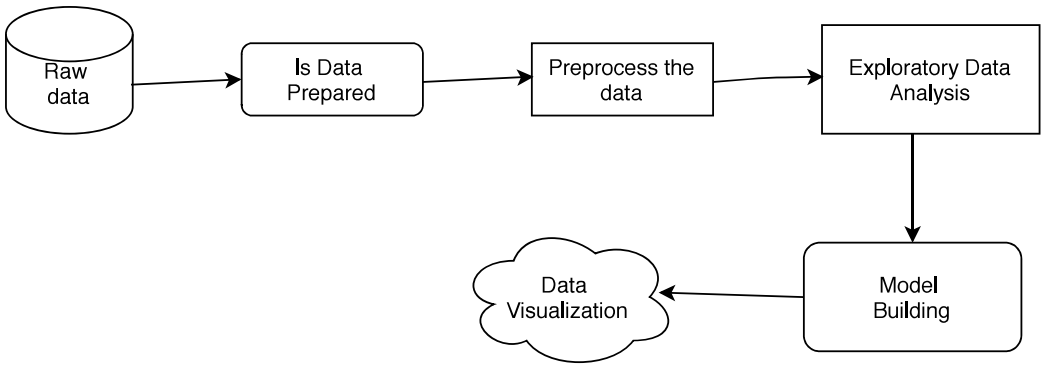

Fig. (2). Data Science Process.

The preface of data analysis is good data exploration and visualization. The EDA ensures the data scientist two-fold:

1. Acquiring more familiarity about the data at hand before developing appropriate business models.
2. Validating the right questions without skewing their assumptions when delivering the results to the business stakeholders.

The following section discusses the methods used in EDA by illustrating the interesting examples wherever suitable.

SUBJECT INDEX